DOCUMENT RESUME

ED 078 061 TM 002 878

AUTHOR Bayuk, Robert J.
TITLE The Effects of Choice Weights and Item Weights on the Reliability and Predictive Validity of Aptitude-Type Tests. Final Report.
INSTITUTION Pennsylvania Univ., Philadelphia.
SPONS AGENCY National Center for Educational Research and Development (DHEW/OE), Washington, D.C. Regional Research Program.
BUREAU NO BR-1-C-047
PUB DATE Mar 73
GRANT OEG-3-71-0108
NOTE 77p.

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS *Aptitude Tests; Correlation; *Predictive Validity; Response Style (Tests); Scoring Formulas; Statistical Analysis; Technical Reports; Test Interpretation; *Test Reliability; Test Results; *Weighted Scores

ABSTRACT

An investigation was conducted to determine the effects of response-category weighting and item weighting on reliability and predictive validity. Response-category weighting refers to scoring in which, for each category (including omit and "not read"), a weight is assigned that is proportional to the mean criterion score of examinees selecting that category. Item weighting refers to the application of multiple regression techniques to maximize the relationship between a composite of item scores and a criterion. The study of the effects of weighting on reliability indicated that scores resulting from response-category weighting were significantly more reliable than scores corrected for chance success. Response-category weighting in concert with item weighting resulted in scores significantly less reliable than scores corrected for chance success. The study of the effects of the weighting on predictive validity indicated that no gain in predictive validity accrued through the use of response-category weighting as opposed to scores corrected for chance success. Response-category weighting with item weighting resulted in scores significantly more reliable than scores corrected for chance success. (Author/CK)

FILMED FROM BEST AVAILABLE COPY

ED 078061

FINAL REPORT

PROJECT NO. 1-C-047
CONTRACT NO. OEG-3-71-0108

ROBERT J. BAYUK, JR.
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

The Effects of Choice Weights and Item Weights on the Reliability and Predictive Validity of Aptitude-Type Tests

March 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION, AND WELFARE

Office of Education

National Center for Educational Research and Development
(Regional Research Program)

ED 078061

TM 002 878

Final Report

Project No. 1-C-047
Contract No. OEG-3-71-0108

THE EFFECTS OF CHOICE WEIGHTS AND ITEM WEIGHTS ON THE RELIABILITY AND PREDICTIVE VALIDITY OF APTITUDE-TYPE TESTS

Robert J. Bayuk, Jr.

University of Pennsylvania

Philadelphia, Pennsylvania 19104

March 1973

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
National Center for Educational Research and Development

ABSTRACT

The primary objective of the investigation was to determine the effects of response-category weighting and item weighting on reliability and predictive validity. Response-category weighting refers to scoring in which, for each category (including omit and "not read"), a weight is assigned that is proportional to the mean criterion score of examinees selecting that category. Item weighting refers to the application of multiple regression techniques to maximize the relationship between a composite of item scores and a criterion.

The study of the effects of weighting on reliability indicated that scores resulting from response-category weighting were significantly more reliable than scores corrected for chance success. Response-category weighting in concert with item weighting resulted in scores significantly less reliable than scores corrected for chance success.

The study of the effects of weighting on predictive validity indicated that no gain in predictive validity accrued through the use of response-category weighting as opposed to scores corrected for chance success. Response-category weighting with item weighting resulted in scores significantly more reliable than scores corrected for chance success. Further research is necessary to refine the application of response-category and item weighting to clarify interpretation of obtained weights.

ACKNOWLEDGMENTS

The writer is grateful to Dr. Frederick B. Davis, University of Pennsylvania, who provided indispensible guidance and assistance in every phase of this investigation.

The writer is indebted to Dr. Ralph C. Preston, Director of the Reading Clinic, Graduate School of Education, University of Pennsylvania, and members of his staff for providing portions of the data used in this investigation.

This investigation was made possible by a grant from the Regional Research Program of the National Center for Educational Research and Development (Region III).

Robert J. Bayuk, Jr.
Principal Investigator

University of Pennsylvania
Philadelphia, Pennsylvania

CONTENTS

Abstract..... ii

Acknowledgments..... iii

List of Tables..... vi

Chapter I: PROBLEMS AND OBJECTIVES..... 1

Weighting Item Scores..... 2

Weighting Item Response Categories..... 4

Chapter II: REVIEW OF THE LITERATURE..... 6

Weighting Test Items..... 6

Differential Weighting of Item Response-Categories..... 9

Summary..... 20

Chapter III: THE RELIABILITY STUDY..... 21

Purpose..... 21

Tests Used..... 21

Samples..... 21

Scores to be Compared..... 22

Determination of Scoring Weights for Method W3..... 25

Determination of Scoring Weights for Method W4..... 25

Estimation of Parallel-Forms Reliability Coefficients for Total Scores on Tests C and D Obtained by Four Different Scoring Methods..... 34

Tests of Significance of Planned Comparisons..... 42

Chapter IV: THE PREDICTIVE VALIDITY STUDY..... 45

Purpose..... 45

Test Used..... 45

Samples..... 45

Scores to be Compared..... 46

Determination of Scoring Weights for Method W3..... 48

Determination of Scoring Weights for Method W4..... 48

Estimation of Predictive Validity Coefficients for the Davis Reading Test Total Scores Obtained by Four Different Scoring Methods..... 55

Tests of Significance of Planned Comparisons..... 58

Chapter V: SUMMARY, DISCUSSION, AND CONCLUSIONS..... 60

Summary of the Reliability Study..... 60

Summary of the Predictive Validity Study..... 62

Discussion and Conclusions of the Reliability Study..... 63

Discussion and Conclusions of the Predictive Validity Study..... 64

References..... 66

LIST OF TABLES

Number	Title	Page
1	Numbers of Examinees in Validation and Cross-Validation Samples 1R, 2R, and 3R for the Reliability Study	23
2	Descriptive Statistics on the Criterion Variables for ALL Samples Reliability Study	24
3	Response-Category Weights for Each of the 96 Items in Form C of the Experimental Reading Test, Sample 1R (N=330)	26
4	Frequency of Response to Each Response Category in Form C of the Experimental Reading Test, Sample 1R (N=330)	28
5	Response-Category Weights for Each of the 96 Items in Form D of the Experimental Reading Test, Sample 1R (N=331)	30
6	Frequency of Response to Each Response Category in Form D of the Experimental Reading Test, Sample 1R (N=331)	32
7	Partial Regression Coefficient for Each Variable in Form C When the Criterion is Normalized Standard Scores on Form D, Sample 2R-C (N=331)	35

Number	Title	Page
8	Partial Regression Coefficient for Each Variable in Form D When the Criterion is Normalized Standard Scores on Form C Sample 2R-D (N=328)	37
9	Multiple Correlation and Significance-Test Summary for the Regression of Normalized Form D Standard Scores on Form C Items Sample 2R-C	39
10	Multiple Correlation and Significance-Test Summary for the Regression of Normalized Form C Standard Scores on Form D Items Sample 2R-D	40
11	Intercorrelations, Means, and Standard Deviations of Several Total Scores on Tests C and D Obtained by Four Scoring Methods in Sample 3R (N=360)	41
12	Descriptive Statistics for Grade-Point Averages for Three Samples of University Freshman	47
13	Response-Category Weights for the Davis Reading Test, Series 1, Form D Sample 1V (N=953)	49
14	Frequency of Response to Each Response Category in the Davis Reading Test, Series 1, Form D Sample 1V (N=953)	51
15	Partial Regression Coefficients for Scores of 80 Items in the Davis Reading Test, Series 1, Form D, for Predicting Freshman Grade-Point Averages Sample 2V (N=953)	53

Number	Title	Page
16	Multiple R for Regression of GPA on the 80-Item Davis Reading Test, Series 1, Form D Sample 2V	56
17	Product-Moment Intercorrelations Among Grade-Point Averages and Davis Reading Test Scores Obtained by Four Scoring Methods in Sample 3V (N=953)	57

CHAPTER I

PROBLEMS AND OBJECTIVES

When the reliability and predictive validity of a test are considered, the effects of examinee motivation, administrative circumstances, and scoring procedures are often neglected when, in fact, they should not be. The investigator generally wants to determine as reliably as possible the rank ordering of a group of examinees on the composite of traits measured by a defined criterion variable. If the investigator is dissatisfied with the test's reliability or predictive validity, or both, several alternatives for improving these characteristics present themselves. Among other strategies, he may replace or revise some of the test items, he may improve the criterion measure with which the test scores are correlated, or he may score the test in a different manner. If the investigator already has a test made up of satisfactory items and a set of criterion scores that are both reliable and unbiased, he may still rescore the test with the hope of improving its efficiency. One scoring procedure that may be employed uses differential choice weights. The problem of differential weighting of only the correct responses in test items or of all choices are usually considered separately. The weighting of these two entities usually can be classified into variable-weighting and fixed-weighting methods.

In variable-weighting methods there is no weight, constant over subjects, applied to a single item or item choice. In these methods each examinee provides subjective probability estimates of how confident he is in making a choice. For example, DeFinetti (1965) proposed that an examinee's store of "partial information" be estimated in terms of a subjective probability made by the examinee to indicate the likelihood that a choice that he has marked as correct is, in fact, correct. Scoring items on this basis may, however, introduce the dimension of willingness to gamble on the part of the examinee (Swineford, 1941). After being trained in the test-taking procedures, the examinee realizes that he can get more credit for marking an item correctly by indicating that he is sure of the correctness of his action than by indicating some lack of confidence in his decision. This procedure introduces an unintended variable into the scores so that the test may no longer measure the trait that it was designed to measure. Other limitations or shortcomings of these methods include the need for multiple responses per item, multiple scoring of answer sheets, and the examinee's difficulty in understanding how to take the test.

Fixed weights, usually derived by multiple-regression procedures, refer to weights for application to all item choices. These are identical for all choices in a given item and are constant for all examinees. Some research workers have suggested that fixed-weighting procedures

have maximum value when only small numbers of items are to be weighted. Fixed weights for each item choice are most commonly used when there is no correct choice; e.g., in personality and interest inventories. For each choice, a fixed weight is generally derived on the basis of the correlation between marking or not marking that choice and some criterion variable; e.g., performance on a job, or membership in one of several defined groups.

Although differential weighting of test items, item choices, or some combination thereof should, in theory, provide gains in test reliability and predictive validity, in practice only small gains generally result. It is this result that has led some psychometricians to conclude that differential weighting is not worthwhile (e.g., Guilford, 1954; Gulliksen, 1950). On the other hand, some investigators (Davis, 1959; Hendrickson, 1971; Reilly & Jackson, 1972) have reported significantly improved reliability coefficients by using weights for each choice in every item.

The objective of the present study is to compare the reliability and predictive validity of test scores when the scoring procedure is based on:

1. a-priori weights of 1 for each correct response and 0 for each incorrect response or omission;
2. a-priori weights of 1 for each correct response, -1/k-1 for each incorrect response, and 0 for omission. This is the conventional procedure for correcting for chance success;
3. cross-validated weights for every item response-category;
4. cross-validated weights for every item response-category after the weights have been adjusted by means of cross-validated partial regression coefficients for predicting a defined criterion.

Weighting Item Scores

The reliability coefficient of a test, t, when all variables are expressed in standard-score form, may be written as:

$$r_{tt'} = \frac{\sum_{i=1}^{n} w_i^2 r_{ii'} + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j r_{ij}}{\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j r_{ij}} . \qquad (i \neq j)$$

Weighting the items of a test may affect the sample test reliability coefficient to the extent that the more reliable items are weighted more heavily than the less reliable items. Kelley has shown (1947, pp. 423-424) that $r_{tt'}$ can be maximized if the item scores

are weighted by the inverse of their variance errors of measurement. For an item, i, the weight w_i may be written as:

$$w_i = \frac{1}{\sigma_i^2 (1-r_{ii'})}$$

In practice, as a single dichotomously scored item varies from 50-per-cent difficulty level in a sample, its variance and its reliability coefficient decrease, thus keeping its variance error of measurement fairly constant in value until the item approaches 0 or 100 per cent in difficulty. At either of these limiting values the item no longer differentiates among examinees in the sample tested; it is not differentiating one examinee from one another and has a variance and a variance error of measurement of zero. As a consequence of the fact that the weights for items that are capable of maximizing the reliability coefficient of the test tend to remain the same for most items of the usual difficulty levels, it makes little difference with respect to test reliability whether the optimal weights are used or are not used. A number of empirical studies have confirmed the conclusions of the analytic formulation of the problem given above. These studies are summarized in the chapter that presents a review of the literature.

In the general case, the correlation of a weighted sum (ws) with an independent variable (c) is:

$$r_{(c)(ws)} = \frac{\sum r_{ci} w_i \sigma_i}{\sqrt{\sum w_i^2 \sigma_i^2 + 2\sum r_{ij} w_i w_j \sigma_i \sigma_j}} \qquad (i > j = 1).$$

More specifically, this equation can be considered to yield a predictive validity coefficient of a test composed of i items with some criterion c. To maximize the relationship $R_{(c)(ws)}$ (the validity coefficient), the proper weights (w_i) are the multiple regression coefficients (beta weights) ($\beta_1, \beta_2, \ldots, \beta_i$) for each item in the test being weighted. The extent to which the multiple-correlation coefficient will exceed the zero-order correlation of the unweighted sum of the test items with the criterion (after cross-validation) depends largely on the degree to which the items differ with respect to their correlations with the criterion variable and with each other. If the items in the test are homogeneous in content, the use of multiple-regression weights is not likely to result in an appreciable gain in test validity. On the other hand, if the test items are heterogeneous (as they are in some cases because they are components of a test that properly measures a complex function), the multiple correlation coefficient might be considerably higher than the zero-order coefficient

$r_{(c)(s)}$. Empirical studies bearing on this point are discussed in the chapter that presents a review of the literature.

Weighting Item Response Categories

If differential weights are assigned to each response category in a multiple-choice item, the number of score categories may be increased beyond the dichotomy of "passing" or "failing" the item. For example, with 5-choice items in which each choice has a different weight, an examinee may receive any one of five different item scores by marking one of the five choices. However, two other response categories are available to him; he may read the item and choose to refrain from marking an answer to it or he may work at a rate slow enough so that he does not have time to read a given item in the time limit. Since scoring weights can be assigned to these response categories, an examinee may obtain any one of seven scores for a 5-choice item.

Guttman showed (1941) that the correlation ratio between a set of scores on one item (when these scores take the form of numerical values assigned to the item response categories) and a set of criterion scores can be maximized by assigning to each item response category a value proportional to the mean criterion score of the examinees who fall in that category. This general least-squares mathematical model for obtaining weights that maximize internal consistency falls under the general heading of scaling. Torgerson (1958) has provided a comprehensive review of these techniques, including Guttman's method, which he categorizes as a method of scaling principal components.

In the present study, Guttman's procedure has been generalized from its application to questionnaires to obtaining weights for all response categories available to examinees who take aptitude and achievement tests. This involves having a scoring weight for each choice in a multiple-choice item, a scoring weight for reading each item and refraining from marking an answer to it, and a scoring weight for not reading the item during the time limit. By including the last two response categories, the scoring system is able to take partially into account such components as personality factors, test-taking strategies, and rate-of-work determinants.

Guttman (1941) outlined an analytical procedure for obtaining the "best" set of numerical weights for each choice in a series of multiple-choice items in the sense that the choice weights would yield the maximum correlation ratio between the sum of weighted item scores and the criterion variable.

The main consideration of the present investigation is the application of Guttman's scaling method to multiple-choice items that, unlike the items considered by Guttman, have a keyed "correct" answer or response. The effects of this scaling procedure, applied to aptitude-

or achievement-test items, can be viewed in terms of the changes in the test's reliability and predictive validity.

Concern over the question of the information carried in the choice among wrong responses in a given test item is evidenced in the literature. In a paper by Powell (1968), the question of the functional role of wrong answers in multiple-choice tests was the main concern. Powell was particularly interested in the amount of potentially useful information that is lost when all distracters of an item are considered in the general category of "wrong responses." Powell, like Davis (1959), observed "...much time is spent... in the preparation of foils for multiple-choice tests. And a proportionally large amount of time is spent by the examinee in making his selection decisions among the alternatives (p. 403)." From these observations, Powell conjectured that the "wrong" answers may indeed have as much discriminating power as the "right" answers.

The present study employs an item response-category weighting method that is a modification of the method originally proposed by Guttman (1941) and is concerned with the effects of item response-category weighting on the reliability and predictive validity of reading tests that measure largely verbal aptitude. The value of the response-category weighting methods described herein is judged in terms of practical as well as statistical significance.

CHAPTER II

REVIEW OF THE LITERATURE

Literature pertaining to two applications of fixed weighting procedures is presented in this review. The first application deals with uniform weighting of test items by applying the same weight to all response categories for the item. The second deals with the differential weighting of response categories for an item.

Weighting Test Items

In general, when a uniform weight is applied to all response categories in an item, the items themselves are usually scored in a conventional manner. That is, the items are usually scored "pass" or "fail," with a score of 1 being applied in the former case and a score of 0 being applied in the latter, by the application of the correction-for-guessing formula, a 1 being assigned to a correct choice, a negative score $-1/(k-1)$ being applied to an incorrect choice, and a 0 being applied to an omitted item. Ordinarily the total test score for an examinee is obtained by summing the item scores over all items in the test.

The numerous empirical studies reporting the use of uniform weighting of all response categories in an item provide fairly overwhelming evidence that it is not effective in increasing the reliability of a test. From formulas presented by Wilks (1938) and Gulliksen (1950) on the correlation of weighted sums it is generally agreed that when the number of predictor variables (items) is large and only positive weights are used, the effects of any weighting system are limited. Even when random sets of positive weights are used the resulting correlations between weighted and unweighted scores are high. However, as Stanley and Wang (1968) point out, uniform weighting of item response categories may still be useful for increasing predictive validity.

Douglass and Spencer (1923) investigated the utility of weighting the exercises or items in objective tests. They obtained correlations of .98-.99 between weighted and unweighted scores on four parts of an algebra test given to 25 secondary-school students. They found analogous correlations for the Henmon Latin Test ($r = .98$) and for the Gregory Test of Languages ($r = .99$). All three examples involved the scoring of the same test items in two different ways. The fact that spuriously high correlations might be obtained as the result of correlating errors of measurement was apparently not considered. Although no conclusions were drawn or recommendations made, they did note that the results were in accord with earlier work by Charters (1920). Douglass and Spencer stated that the weighting procedure was time-consuming, tedious, and

increased the possibility of error in test scoring.

Holzinger (1923) found similarly high correlations between weighted and unweighted scores. On a 40-item test of French grammar, a correlation of .99 between weighted and unweighted scores was obtained. Similar results were obtained for an algebra test and an arithmetic test.

West (1924) reported the results of a fairly thorough investigation of the effects of weighting test items on three different tests. In each case the weighting method was the same. Weights for items were a function of the proportion of examinees who incorrectly answered each item. The first study by West compared weighted and unweighted scores on each of five parts plus the total score on two forms of a reading-comprehension test. Only one of the twelve correlations obtained was below .99. The Army Alpha Test (Form 8) was administered to the same group of 45 secondary-school students and the effects of item weighting on six of the eight parts were studied. Correlation coefficients between raw and weighted scores ranged from .940 to .984. West noted that the intercorrelations of the part scores were similar for both types of scoring.

A third test, a collection of 200 analogies, was divided into five measures of 40 analogies each. The tests were designed so that the accumulated scale values for each test would be the same. Each test was administered to the same group of 45 secondary-school students used earlier. Scoring of each test was done in three ways. An unweighted (raw) score, a Pintner Scale score, and a weighted score were obtained for each test. Intercorrelations of the five tests were computed for each scoring procedure. Correlations between each of the 10 pairs of tests scored in each of three ways were computed. The 30 correlation coefficients varied from one scoring method to the other. In fact, West noted that the rank ordering of subjects based on each of the methods were markedly similar.

West concluded that weighting of test items was generally not valuable for purposes of more accurately differentiating the measured abilities of examinees. He did, however, note that some value might be had in weighting items for purposes of scaling and arranging items in a test and then scoring the items in the conventional "raw-score" manner.

Peatman (1930) attempted to determine the value of Clark's Index of Validity as a weight for true-false test items used in determining a subject's relative standing or grade. Data were obtained for 73 college students on six 25-item true-false quizzes and a final 100-item true-false examination. For the six quizzes the correlations between weighted and unweighted scores ranged from .879 to .970. The same correlation for the longer final examination was .955. A "combined score," an average of all quizzes and the final examination, both weighted and unweighted, yielded a correlation of .978. Peatman concluded on the basis of these findings that weighting of true-false items by the method used was not

justified. The high correspondence between the original and weighted scores resulted in few changes in the relative standing of subjects whose grades were determined by these methods.

Corey (1930) had six psychology instructors evaluate 73 items from a psychology examination with respect to each statement's importance for a general knowledge of psychology. Correlations between each instructor's weighted scores and the raw scores, using 100 randomly selected test papers, were obtained. These correlations ranged from .82 to .96. They were interpreted to indicate that weights assigned by all instructors, save one, noticeably affected the relative standings of the students. It was also found that, in the case of one instructor, 49 per cent of the test papers would have been assigned grades at variance with those assigned using the raw-score method. Corey observed that the grades given by competent judges who weight each test item differently will vary considerably from those grades assigned on the basis of raw scores. He concluded that the objectivity of raw-score weighting is spurious because some items are naturally more important than others. No information as to the reliability of the judge's ratings was presented.

Because the conclusions reached by Corey (1930) disagreed with earlier evidence indicating that item weighting makes no difference, Odell (1931) conducted two studies similar in several respects to the earlier investigation by Corey. In the first study, Odell obtained six sets of weights for a 50-item four-choice test. Weights were determined by random assignment of weights to items in three of the six methods. Even when the "random weights" were used, the correlations between weighted and unweighted scores for 62 test papers ranged from .92 to .99. When weighted scores and scores corrected for guessing were correlated, the range of coefficients remained in the range of .98 to .99. In the second study, a 22-item true-false test was used. Weighting for three of the methods was determined by instructors. Correlations between weighted and unweighted scores ranged from .95 to .98. Odell concluded that little is to be gained from weighting items in objective-type examinations, a conclusion at variance with that of Corey.

Neither Corey nor Odell presented evidence of the reliability or validity of either weighted or unweighted scores. Further, no data were presented on the correlations among the sets of weights obtained from the judges. Odell did reveal, however, that some of the judges in Corey's study attached weights of zero to some items.

A study by Potthoff and Barnett (1923) was concerned with the effects of the weighting of test items on the grades of individuals. Eleven methods were used to score a 100-item examination in high-school American history. Ten of the scoring methods were based on ratings by ten history instructors. One weighting method was the equal or unweighted system ordinarily used. Potthoff and Barnett were primarily concerned with the agreement between the weighted and unweighted scoring

methods with regard to the assignment of grades based on test scores. The average agreement between all raters for all grading categories and the grade assigned by the unweighted method was 88 per cent. The authors cautioned the reader that, even when correlations between weighted and unweighted scores are high ($\sim$.96), letter grades may still disagree considerably in some cases, especially in the middle (B-C) range. Potthoff and Barnett concluded that, for practical purposes, the differences between weighted and unweighted scores are generally so small that they can be disregarded and a great deal of labor can be saved by using the conventional, unweighted method of test scoring.

Stalnaker (1938) considered the question of weighting as it affects the essay-type examination question. Citing several examples of weighting various College Entrance Examination Board essay questions, Stalnaker reported correlations between weighted and unweighted scores as being above .97 for tests in a variety of subject areas. Even when weights were assigned to items based upon the position of the item in the test, the obtained correlation between weighted and unweighted scores was .99. This indicated to Stalnaker that, because of the small net effect and the laboriousness of the weighting procedures employed, weighting of items is not extremely valuable.

Although Stalnaker's paper provided no mathematical treatment of the effects of weighting test items, Wilks (1938) demonstrated the effects analytically. Wilks showed that, in a long test (50-100 items), when the item responses are positively intercorrelated, weighting items has little effect on the rank order of scores. In fact, when the number of items is large, the rank order of scores tends to become stable, or invariant, for different methods of obtaining linear scores.

The foregoing review of the empirical studies of the effects of weighting test items leads to the general conclusion that it is not worth the trouble to apply the same weight to all choices in a multiple-choice item or to credit assigned for an essay question. And Wilks' analytical paper provides the mathematical rationale and proof of why this conclusion is warranted. This conclusion must not, however, be applied to the use of differential response-category weights. There is evidence that differential weighting of incorrect responses can be of considerable value for increasing test reliability.

Differential Weighting of Item Response-Categories

Empirical investigations of weighting response categories of test items differentially stems from work using interest and personality inventories. Some of the earlier work using this approach to item scoring was done by Strong (1943) and Kuder (1957). Both of these investigators have reported positive empirical evidence of the value of differentially weighting response categories of items in interest inventories. Their work, however, involved the weighting of response

categories of questionnaire-type items with no correct answer. Weighting response categories of items with no correct answer is generally considered to be scaling and does not directly relate to this study. On the other hand, several different scaling techniques have been shown to be applicable to weighting response categories in aptitude-type tests. Of particular interest is a method proposed by Guttman (1941) for use in scoring interest inventories. Analytical and empirical evidence of the utility of differentially weighting response categories in aptitude and achievement tests is of particular importance to the present study.

One of the earliest studies using a weighted-choice test-scoring procedure with an ability-type test was conducted by Staffelbach (1930). Using a sample of 244 eighth-grade students for whom both test data and criterion data (semester grade averages) were available, Staffelbach obtained raw-score regression coefficients for three scores on a 60-item true-false test; number right, number wrong, and number omitted. The regression coefficients were .5017, -.5489, and .3559 for the rights, wrongs, and omits, respectively. Wrong responses were weighted slightly more heavily in the negative direction than were the right responses in the positive direction. Omits were assigned a positive weight. Thus, marking the correct response and recognizing inability to answer were both given positive weights in this system.

Since the Staffelbach study involved a true-false test the differential weighting was not of incorrect responses but of incorrect as opposed to omitted responses. In this sense the weighting is similar to the now-common correction-for-guessing formula. In fact, the weights for right and wrong responses are quite similar in that they are approximately equal, but differ in sign.

Kelley (1934) described a response-category weighting procedure that takes into account the item-criterion correlation when both variables are dichotomous. The formula presented by Kelley is

$$W = b_{21}/\sigma^2_{b_{21}}$$

where W = the response weight;

b_{21} = the regression coefficient of the criterion on the item, and;

$\sigma^2_{b_{21}}$ = the variance of the regression coefficient.

This procedure for weighting item choices or, actually, any responses that are dichotomous, was recommended by Kelley for use with interest-inventory items like those developed by Strong (1943)

Guilford, Lovell, and Williams (1942) investigated the effects of differential response-category weighting on test reliability and

predictive validity. The items for which response-category weights were obtained consisted of the first 100 (the first 101 minus one item known to be defective) items of a 308-item final examination in general psychology. A total test score was obtained by scoring all 307 items with a correction for chance success. The directions to the examinees did not state this fact, however. From 300 answer sheets drawn at random data from the 100 sheets having the highest total scores and from the 100 sheets having the lowest total scores were used to obtain approximations to the per cent of the sample marking each category and to the phi coefficient between total test score (treated as a dichotomy) and the dichotomy of "mark" or "not mark" each response. These data provided the basis for the response-category weights as described by Guilford in an earlier study (1941).

Reliability coefficients of scores based on weighted response categories and on the conventional scoring formula were obtained from a sample of 100 papers drawn from the 300 used to establish the category weights. Scores on odd and even items were obtained by both scoring procedures and the correlations of odd and even scores were corrected by the Spearman-Brown formula. The reliability coefficients for scores derived from the 100 items were .922 for the weighted scores and .899 for the unweighted scores. For the scores derived from the first 50 items, the analogous reliability coefficients were .860 and .844. Similar reliability coefficients for the first 20 items were .677 and .649. The statistical significance of the difference between each pair of reliability coefficients could not be tested or estimated without additional data. Thus, no conclusions about the statistical significance of the differences were reached.

Any comparison of the difference between the reliability coefficients in each pair must take into account the fact that the 100 answer sheets used to compute the reliability coefficients for the weighted scores were drawn from the same sample on which the weights were established. That this procedure leads to spuriously high reliability coefficients must be considered a serious possibility. Even with this in mind the data suggest that the use of response-category weights of the type used by these investigators provided scores little more reliable than those obtained through conventional scoring procedures.

It is quite possible that the items themselves were of a nature that did not encourage the use of partial information for marking choices among distracters. Also, the items may have been easy, thus making the use of differential response-category weights less likely to contribute reliable information to the test scores.

Several investigators (Coombs, Milholland, & Womer, 1956; Dressel & Schmid, 1953; Hawver, 1969) have presented scoring procedures that attempt to assess partial knowledge available to an examinee.

The Dressel and Schmid study (1953) was among the first to investigate modified multiple-choice items to determine whether they could be made to be more discriminating. Five groups of approximately 90 college students each first received a "standard test." This standard test was used to determine the equality of the groups. Three of the groups then took a single 44-item multiple-choice test but with differing instructions on how to respond to each item. The first group received instructions that the score was to be number right. The second group received instructions to mark as many choices per item as necessary to insure marking the correct choice. This "free-choice test" was believed to take the student's certainty of response into account. A third group was asked to indicate certainty of response to each item by assigning a number from a 4-point "certainty scale." This was termed the "degree-of-certainty test." A fourth group took a modified version of the 44-item test with the choices in each item changed so that more than one choice could be correct. The students were informed of this fact. This "multiple-answer test" was designed to compel the students to assess each item more thoroughly. Finally, the fifth group took a modified version of the 44-item test with the choices changed so that there were two correct choices per item. Examinees were informed that the scores would equal the number of items marked correctly.

Comparing the reliabilities and validities of the tests on which the five special scoring methods were used with those of the standard test, Dressel and Schmid reported no significant differences.

Coombs, Milholland, and Womer (1956) presented reliability coefficients of three 40-item tests that had been administered and scored conventionally and in such a way as to incorporate the effect of using partial information in marking test items. The reliability coefficients for the conventional and special procedures, respectively, were .72 and .73 for a vocabulary test, .64 and .70 for a driver-information test, and .89 and .91 for an object-aperture test. The statistical significance of the difference between reliability coefficients in each of the pairs of coefficients could not be obtained since the coefficients were obtained by Kuder-Richardson formula no. 20.

The authors provided data showing that the examinees used partial information in answering items in the vocabulary and driver-information tests. Their analysis of responses of examinees to difficult and easy items provides a statistically significant confirmation of the expectation that the reliability of a test composed of difficult items is more likely to be increased by the use of response-category weights in scoring than the reliability of a test made up of easy items.

Nedelsky (1954a) described a method by which the choices in multiple-choice items could be classified into three general categories. Instructors classified responses as R responses or right answers,

F responses or wrong responses that would have appeal only to the poorest students, and W responses, wrong responses other than F responses. Another paper by Nedelsky (1954b) was concerned with the uses of the F score made by an examinee and the number of F responses chosen in a multiple-choice test. The properties of the F score were studied alone as well as in combination with the R score. The composite (C) score resulting from this combination superficially resembles the common "formula-scoring" procedure that provides a penalty for guessing. The score C is defined as:

$$C = R - F/f$$

where C is the composite score;
R is the "rights" score;
F is the F score (number of F responses chosen), and;
f is the average number of F responses per item in the test.

In this study, Nedelsky analyzed a 113-item multiple-choice test given to 306 students completing a course in the physical sciences. Grades for the students were determined on a basis of the R scores. The "experimental group" contained all students receiving a grade of D or F and a representative sample of those who received higher grades. Kuder-Richardson reliability estimates were calculated for R, F, and C scores for the A, B, and C students, for the D and F students, and for the total group. These coefficients indicated that the R score had a negative reliability for the D and F students. The F-score reliability of .42 was the highest obtained for this group, the C-score reliability being .26. Interestingly, the C-score reliability calculated for the A, B, and C group and the total group exceeded the R-score reliability by at least .02 in the first case and .03 in the second. It was noted, however, that only 70 of the 113 items in the test had any F responses in them.

Over-all, the C score was considered to be the most reliable score calculated from the data on this sample of examinees. Nedelsky posits that the F score "...furnishes evidence of the existence of an identifiable ability to avoid gross error in a given field and for considerable differences in this ability among the poorest students of a class (p. 464)."

Merwin (1959) provided a detailed theoretical analysis of six methods of scoring three-choice items. Methods using two, three, and six response patterns were considered in conjunction with consecutive integer weights and weights which maximize the correlation of item scores with the criterion. If each subject is instructed to rank the three choices in an item according to their attractiveness, there are only six different response patterns available. Thus, six different scores can be assigned to a single item. For example, the permuted response patterns of "abc," "acb," and "cab," etc. are assigned different weights. In the three-score paradigm, only the rank of the correct alternative is considered. And in the two-score scheme, the subjects

merely indicate their first choice. This third method is the common "rights-only" method of scoring.

Weighting of response patterns was accomplished through either integer weights or weights proportional to the mean criterion score for subjects choosing the particular response pattern. The weights in the latter case were identical in kind to those described by Guttman (1941). Each scoring and weighting combination was studied by systematically varying the item parameters and studying the effects on the "efficiency indexes." What Merwin termed the "efficiency index" is actually the product-moment correlation coefficient between item scores and a specified criterion. Merwin summarized his theoretical study by saying that the use of the six-score scheme, in combination with the Guttman-type weights, will always yield item validity efficiency as high or higher than any other method studied. Merwin also pointed out, however, that the increases are relatively small and would be smaller after cross-validating the obtained response weights. For efficiency and ease of scoring, the "best" method studied was that using three integer weights, +1, 0, and -1 with the three-score scheme.

The two papers that follow are considered in much greater detail than others included in this review because of their direct relation to the present investigation. The article by Davis and Fifer (1959) presents empirical evidence of the value of response-category weighting of the kind used in the present study. The second article by Davis (1959) describes analytically choice-weighting procedures that he recommends.

Davis and Fifer (1959) investigated the effects of response-category weighting of multiple-choice items on the reliability and validity of an achievement test. From approximately 300 arithmetic-reasoning items constructed especially for this study, two matched sets of 45 items were chosen. In addition, two matched sets of 5 items testing computational skills were also constructed and included in the arithmetic-reasoning tests. These "computational" items, when scored appropriately, served to cancel some variance in the test scores that might be attributed to computational facility and not arithmetic reasoning.

Two mathematicians, working independently, assigned weights to each choice in the two 45-item tests. These weights were on a seven-point scale from -3 to +3. These two sets of weights were then reconciled to obtain one set of a-priori weights for all choices in the two tests (5022 and 5023 were the test labels). This same procedure was carried out for the two sets of five "computational" items. The signs of the weights for the "computational" items were reversed, however, to make them serve as a "suppressor" variable for computational facility.

Both tests (5022 and 5023) were administered to a sample of over 1000 airmen at Lackland Air Force Base. From this initial group, answer sheets of a subsample of 370 airmen were drawn at random and scored using the a-priori weights. Empirical weights, expressed as biserial

correlations between total test score and marking or not marking a choice, were calculated for each choice. The empirical weights were then modified so that no wrong answer was allowed to have a scoring weight higher than that of the correct answer to the item of which it was a part.

The remainder of the sample from which the 370 cases had been drawn at random was used to test the effect of these differential response-category weights on the reliability and validity of the two tests. Four scores were obtained for each examinee in this sample. They were: 1) number correct on test 5022; 2) number correct on test 5023; 3) the sum of the choice weights for choices marked on test 5022, and; 4) the sum of the choice weights for the choices marked on test 5023.

After the raw scores had been converted to normalized standard scores, a parallel-forms reliability coefficient for the unweighted scores on tests 5022 and 5023 was calculated by correlating the "number-rights" scores on these tests. The obtained coefficient was .6836. By correlating the empirically modified weighted test scores for forms 5022 and 5023, a parallel-forms reliability coefficient of .7632 was obtained. After these r's had been converted to Fisher's z values, the difference in z's was found to be statistically significant ($p < .001$). Davis and Fifer noted that this increase in reliability would have been obtained if the tests had been scored "number right" only after their lengths had been increased by 50 per cent.

Two criterion measures were used in assessing increases in validity due to choice weighting of these two tests. One criterion consisted of teachers' ratings of pupil's abilities to solve arithmetic-reasoning problems. The second consisted of scores on a free-response version of items in either 5022 or 5023. A sample of 251 high-school students was divided into four groups. Each group received a free-response version of either 5022 or 5023 and a multiple-choice version of 5022 or 5023. Administration of the different forms was counter-balanced in the four groups to guard against testing-sequence effects. The two groups receiving the multiple-choice version of 5022 were combined, as were the two groups receiving the multiple-choice version of 5023. Validity coefficients were obtained between the multiple-choice tests (scored by the two methods), teacher ratings, and the free-response versions of 5022 and 5023. The two coefficients between the teachers' ratings and the multiple-choice tests scored "rights only" and by empirical weights were .39 and .42, respectively. The coefficients between the multiple-choice tests scored both ways and the free-response test were .69 and .68, respectively. Neither of these differences approached statistical significance. Davis and Fifer concluded that significant increases in test reliability can be gained without reducing test validity by using weights for each choice of a well-constructed test.

Davis (1959) is more explicit about a method of estimating

choice weights that he recommends for practical use. The procedure for obtaining choice-weights that tends to maximize the correlation of any set of items with any given criteria is quite similar to that described by Guttman (1941) Guttman's procedure entails the calculation of the mean criterion score for the group of examinees that select each choice in every multiple-choice item. The actual choice weights are proportional to these mean criterion scores. As Davis pointed out, this procedure would be extremely laborious without the use of high-speed computers. An alternative, short-cut procedure suggested by Flanagan (1939) was used by Davis. This method provides approximations to the Guttman weights by simply reading them from a table published by Davis (1966).

The Flanagan-Davis procedure entails the estimation of $_{ik}\bar{z}_c$, the weight for choice k of item i. The symbol $_{ik}\bar{z}_c$ denotes the mean criterion standard score for the group of examinees who marked choice k of item i. To estimate this weight the correlation $r_{z_{ik}z_c}$ between the item-choice standard scores, z_{ik}, and the criterion standard scores, z_c, can be read from a table devised by Flanagan (Flanagan and Davis, 1950) if the per cents of examinees in the upper and in the lower 27% of the criterion distribution who selected the choice are known. Since $\bar{z}_{ik}$ for item i can be read as the normal deviate corresponding to p_{ik}, the per cent who mark the choice, then $_{ik}\bar{z}_c$ can be estimated from the regression equation. Davis (1966) provided a table for this purpose.

Davis determined the accuracy of this estimation procedure by actually calculating the mean criterion standard scores for examinees responding to each item choice in a 45-item arithmetic-reasoning test. The estimation procedure was also carried out for each item choice in the same test. The obtained correlation between computed means and estimated means for the 45 correct choices was .94. For the 180 distracters the correlation was .91. These correlations and the close similarity of the means and standard deviations of the sets of weights showed that the estimation procedure is highly satisfactory.

To assess the reliability of weights estimated by this procedure, Davis obtained two samples of 370 examinees who took tests 5022 and 5023, the parallel-forms arithmetic-reasoning test used in Davis and Fifer's investigation (1959). Choice weights for both tests were estimated for the two samples. A correlation between the weights for the two tests estimated in two independent samples constituted a reliability coefficient for the weights. Davis found the reliability coefficient of the correct response weights to be .64 and, of the distracters to be .67. These coefficients are significantly different from zero, are moderately high, and could be increased by using larger samples for establishing the weights.

More recently, Sabers and White (1969) reported an empirical study of the scoring procedure previously described by Davis and Fifer (1959) and Davis (1959). These investigators used four groups of

examinees, two groups enrolled in a modern mathematics program and two groups enrolled in a traditional algebra course. All choices on the Iowa Algebra Test were weighted using a chart devised for that purpose by Davis (1966). The criterion measures were 40-item multiple-choice tests scored number correct. Sabers and White cross-validated the weights by applying the weights derived on one group to the other group in the same mathematics category. Non-significant increases in reliability and validity were reported.

The main focus of an investigation by Hendrickson (1971) was to determine the effects of choice (response-category) weighting on the internal-consistency reliability of four subtests of the Scholastic Aptitude Test (SAT). The effects of the weighting scheme on the intercorrelations of the subtests and the regression of scores derived from Guttman weighting on those obtained through the conventional formula-scoring method were also investigated.

The first study by Hendrickson compared the internal consistency reliability coefficients of four subtests of the SAT when they were scored with the conventional correction for chance success and with cross-validated Guttman weights. Comparisons for male and female examinees were treated separately to ascertain any sex-related differences in the effects of choice weighting.

The effective increase in test length varied from subtest to subtest and between sex groups but was no less than 19%. That is to say, a subtest could be reduced in length by almost 20% and, if scored using Guttman weights, would have the same internal consistency as the longer test. Overall, the average effective increase in test efficiency was 49%. Thus, the use of Guttman weights could save considerable testing time without loss of reliability. As Hendrickson points out, the Guttman weighting scheme depends upon the correctness of the assumptions that (a) the quality of response categories differs, and (b) that groups of similar levels of knowledge about the point being tested tend to choose the same category.

Another part of the investigation revealed a significant linear relationship between Guttman and formula-score distributions. Inspection of the plot of the regression of Guttman scores on formula scores showed greater dispersion of Guttman scores at lower values of formula scores. This was taken to indicate that Guttman weighting affects low-scoring examinees more than high-scoring ones. Nedelsky (1954b) demonstrated a similar effect using another weighting scheme.

A comparison of the response-category weights for men and women indicated that, when the weights derived for each sex were interchanged, the distribution of total scores was essentially unaltered. Hendrickson did, however, indicate that while the sexes did not respond differently to the items as a whole, they did respond differently to the choices. It was suggested that this may be a neglected source of bias in testing procedures that is deserving of attention.

In sum, Hendrickson found that Guttman weighting resulted in improved internal-consistency reliability for certain subtests of the SAT. The effects were more pronounced for the verbal subtests, but the weighting procedure also was beneficial in the quantitative subtests. As expected, a linear relationship was found between scores derived from Guttman weights and those derived through conventional formula scoring methods.

Reilly and Jackson (1972) conducted an investigation quite similar in many ways to the present one. They attempted to provide additional evidence of the value of empirical choice weighting in improving the internal-consistency reliability, parallel-forms reliability and validity of a high-level aptitude test, the Graduate Record Aptitude Examination (GRE).

Three types of scoring procedures were employed. One was the conventional formula scoring. A second involved weighting item-response categories by assigning the mean standard score on the remaining items for all persons marking that choice. This second procedure is essentially the one employed by Hendrickson (1971). The third weighting procedure involved assigning to each option in an item a weight which was the mean standard score on the corresponding parallel-form of all persons choosing that option.

Cross-validated weighting procedures on the sub-forms of the GRE revealed substantial increases in both internal-consistency reliability and parallel-forms reliability. The increases in both types of reliability follow a similar pattern with increases in effective changes in test length ranging from one and one-half times to more than twice the original length for the verbal sub-forms of the GRE.

The effects on improving test validity were less impressive. Using a sample different from those used to obtain the empirical weights, weighted and unweighted GRE scores were used to predict grade-point average (GPA) for over 4,000 college students. The weighted scores produced a multiple R .05 less, on the average, than the conventional formula score. Thus, empirical choice weighting to improve reliability did not lead to improved predictive validity for the GRE verbal or quantitative scores.

Item response-category weighting, when the weights are based upon procedures similar to that described by Guttman (1941), may lead to improved internal-consistency and parallel-forms reliability when the appropriate criterion is employed. This has been shown by Davis and Fifer (1959), Hendrickson (1971), and Reilly and Jackson (1972).

Success in improving predictive validity when a test is weighted to increase reliability has been illusory. Davis and Fifer (1959) obtained no significant change in the validity of a mathematics-reasoning test by using differential choice weights. Reilly and

Jackson (1972) obtained lower validity coefficients with choice-weighted scoring than with the conventional scoring with correction for chance success

The empirically derived weight for the "omit" category for an item has been discussed recently by several authors (Green, 1972; Hendrickson, 1971; Reilly & Jackson, 1972). Although Green admits that reliability can be improved by using Guttman weights, Green presents arguments against the use of such weights for one reason. The Guttman weight for omission of an item usually penalizes the examinee severely. In his investigation, Green found that, in general, people who omit items obtain lower scores on a test than those who guess when in doubt about the correct alternative. Because test directions often caution examinees about guessing, Green is of the opinion that it is unethical to use Guttman weights for scoring.

Hendrickson (1971) suggested that weighting the distracters and omit categories had more effect on scores than weighting the correct category. Like Green (1972), Hendrickson found that examinees who tended to omit items also tended to score lower on the test as a whole than examinees who mark incorrect categories. Gains in internal consistency or parallel-forms reliability seem to be due to the effects of weighting on low-scoring examinees. Since low-scoring examinees tend to mark more distracters and omit more items than high-scoring examinees the effects of Guttman weighting are more strongly felt by those at the low end of the score distribution.

The weights for the omit categories for the GRE test items used by Reilly and Jackson (1972) were not what the investigators expected. Examinees were given a bonus for not responding to some of the verbal items. For the quantitative items examinees always received a penalty for omitting an item. The investigators, like Slakter (1967), suggested that, while the propensity to omit items is reliable, it is not valid for predicting some external criterion. This was offered as an explanation for the decreases in validity in spite of the increases in relaibility.

It may be concluded from the recent work of Hendrickson (1971) and Reilly and Jackson (1972) that increases in reliability can be attributed primarily to the differential weighting of distracter and omit categories. In particular, weighting of the omit category seems to provide these increases because omitting items is a characteristic of certain examinees and the effects of this characteristic are reliable. However, as Green (1972) points out, instructions for multiple-choice examinations where a correction-for-guessing formula is used regularly, caution the examinee about wild guessing. The implication for the examinee is to omit when in doubt. Those examinees who omit items tend to be penalized for following directions. It would seem, then, that either the directions for test taking should be changed or the category weight for omitting an item should penalize the examinee less. It seems that the examinee who is aware of what he does not

know should not be penalized more than the examinee who is not aware of what he does not know and selects incorrect answers.

Summary

It has been found experimentally that weighting the correct responses to some items in a test more than others usually has no appreciable effect on test reliability or validity. The mathematical explanation of this finding has been provided by Wilks (1938).

On the other hand, the differential weighting of all choices in each item in a test can have a marked effect on test reliability. As Davis and Fifer (1959) indicated, the differences among the weights assigned to the incorrect choices in an item mainly account for this effect.

The results of differential response-category weighting on test validity depend on the criterion variable used for establishing the weights. It is possible that a set of weights capable of increasing test reliability may decrease test validity for specified criteria. The extent to which this happens in practice is not yet clear.

CHAPTER III

THE RELIABILITY STUDY

Purpose

The purpose of the reliability study in this investigation of the effects of differential choice weighting on test reliability and validity was to compare the parallel-forms reliability coefficients of Forms C and D of the Davis Experimental Reading Tests (Davis, 1968), when scores were obtained by four different methods.

Tests Used

The nature and development of the Davis Experimental Reading Tests, Forms C and D, were described in detail by Davis (1968). Twelve items testing each of eight basic reading skills comprised each form of the test. Each item was made up of a stem and four choices. For additional information about these tests, the reader is referred to the article cited.

Samples

Davis (1968) administered his Experimental Reading Tests in the fall of 1966 to approximately 1,100 twelfth-grade pupils in academic high schools in the suburbs of Philadelphia. Since the tests were designed to measure several aspects of comprehension in reading, time was allowed for every pupil to try every item at each of two testing sessions and schools drawing largely from middle-class and upper-class homes were used. These procedures minimized the effects of the mechanics of reading on the test scores.

From Davis's basic list of examinees, three groups were drawn at random without replacement. Within the first group, two samples (denoted 1R-C and 1R-D in Table 1) of 330 examinees who took Form C and whose corresponding answer sheets for Form D were identified in the group and 331 examinees who took Form D and whose corresponding answer sheets for Form C were identified in the group.

Within the second group, two samples (denoted 2R-C and 2R-C in Table 1) were formed, consisting of 328 examinees who took Form C and whose corresponding answer sheets for Form D were identified in the group and 331 examinees who took Form D and whose corresponding answer sheets for Form C were identified in the group.

The third group was made up of 360 examinees for whom answer sheets for both Forms C and D were available. This sample is denoted 3R in Table 1. Table 2 provides descriptive statistics pertaining to all of the samples used in the reliability study.

Scores To Be Compared

The four methods for obtaining scores to be used in obtaining parallel-forms reliability coefficients for Tests C and D in Sample 3R are as follows:

W1: For each item, examinees were credited with 1 point for a correct response, 0 for an incorrect response, and 0 for omission (failure to mark any choice as correct after reading the item). The total test score consisted of the sum of the item scores in it. This is commonly called "number-right scoring."

W2: For each item, examinees were credited with 1 point for a correct response, 0 for omission, and $-1/(k-1)$ for an incorrect response (where k represents the number of choices per item). This is commonly called "formula scoring," and embodies a correction for chance success.

W3: For each item, examinees were credited with scores based on weights assigned to each choice and to the response category of omission (failure to mark any choice as correct after reading the item). Each scoring weight was made proportional to the mean criterion score for examinees who fell in a given response category. The criterion scores for establishing scoring weights for Form C were total scores obtained on Form D by method W2 in Sample 1R-C. The criterion scores for establishing scoring weights for Form D were total scores obtained on Form C by method W2 in Sample 1R-D. The total scores obtained by method W3 consisted of the algebraic sum of the scoring weights for the 96 response categories (one per item) selected by each examinee on Form C or Form D.

W4: For each item, the examinees were credited with scores based on modified scoring weights assigned to each choice and to the response category of omission. Each of the scoring weights obtained by method W3 was "modified" by multiplying it by the partial regression coefficient that would maximize the multiple correlation between a set of linear composites of the 96 item scores in Form C (or in Form D) and a set of specified criterion scores. For Form C, the criterion scores consisted of total scores on Form D obtained by method W2 in Sample 2R-C. For Form D, the criterion scores consisted of total scores on Form C obtained by method W2 in Sample 2R-D.

Table 1

Numbers of Examinees in
Validation and Cross-Validation Samples 1R, 2R, and 3R
for the Reliability Study

Sample	Test Form	
	C	D
1R	330	331
2R	328	331
3R	360	360

Table 2

Descriptive Statistics on the Criterion Variables* for All Samples Reliability Study

Statistics	Form C Sample 1R	2R	3R
N	330	328	360
Mean	55.493	55.229	55.202
Variance	453.253	437.391	417.217
St. Dev.	21.290	20.914	20.426
Range	87.670	92.000	93.330
Skewness	- 0.666	- 0.634	- 0.628
Kurtosis	- 0.551	- 0.476	- 0.330

Statistics	Form D 1R	2R	3R
N	331	331	360
Mean	54.727	54.282	54.605
Variance	415.354	366.103	336.061
St. Dev.	20.961	19.134	18.332
Range	84.000	90.340	89.330
Skewness	- 0.445	- 0.499	- 0.410
Kurtosis	- 0.747	- 0.594	- 0.331

*Note.--The criterion variable differed for the groups. Depending upon the group the criterion was either Form C score or Form D score.

Determination of Scoring Weights for Method W3

Guttman (1941) showed that the correlation ratio between item scores and a set of criterion scores could be maximized by scoring an item with a weight for each choice proportional to the mean criterion score of examinees who marked that choice. In this study, his procedure has been broadened from use with questionnaires to use with multiple-choice items of any kind and from its use to obtain scoring weights for item choices to use for obtaining scoring weights for other response categories, such as omission (failure to mark any choice as correct after reading the item) or failure to mark any choice because lack of time did not permit reading the item.

To obtain W3 scoring weights for each of the possible five response categories for each item in Test C, the answer sheets for 330 examinees who made up Sample 1R-C were used. Their raw scores (after correction for chance success) on Test D were obtained. These corrected raw scores were then converted to normalized standard scores with a mean of 50.000 and a standard deviation of 21.066. These served as criterion total scores.

Next, the mean criterion total score on Form D of those examinees who fell in each response category for each of the 96 items in Test C was calculated. These means were then transformed linearly so that, within each item, the sum of the products of each transformed mean and the number of examinees entering its calculation was made equal to zero. This constraint was suggested by Guttman (1941). The transformed mean criterion score for each item response category was used as the weight in method W3.

Analogous scoring weights were then obtained for each of the 96 items in Form D by using Sample 1R-D. The W3 response-category weights for Test C are shown in Table 3 and the numbers of examinees on which they are based are shown in Table 4. Analogous data for Test D are shown in Tables 5 and 6.

It should be noted that these scoring weights based on Samples 1R-C and 1R-D were free from spurious inflation because the criterion scores for the weights established for Test C came from Test D and the criterion scores for the weights established for Test D came from Test C.

Determination of Scoring Weights for Method W4

It is well known that the best linear combination of variables for predicting a specified criterion variable can be obtained by using partial regression coefficients to weight each predictor variable. The method used to obtain W4 weights in this study treats each of the 96 items in Test C, scored by W3 weights established in Sample 1R-C, as a variable for predicting total scores on Test D obtained by applying the

Table 3

Response-Category Weights for Each of the 96 Items in Form C of the Experimental Reading Test, Sample 1R (N = 330)

Item	Response Category				
	A	B	C	D	Omit
1	-1.14205	0.06480	-0.00787	-1.56594	-0.49780
2	-0.74449	-0.32822	0.18046	-0.62713	0.23030
3	-0.79691	0.09905	0.28234	-0.49471	-0.60765
4	-0.27261	-0.47775	0.47739	-0.87487	-0.13875
5	-0.69704	-1.11047	-0.38440	0.22828	0.06853
6	0.38060	-0.39871	-0.74398	-0.55657	0.03911
7	-0.47625	-0.53673	-0.52391	0.51634	0.05381
8	0.37707	-0.07124	-0.67509	-0.51037	-0.16048
9	-0.20742	0.50107	-0.26983	-0.71423	-0.90074
10	-0.29144	-0.88729	-0.51215	0.57248	-0.03581
11	-0.85409	-0.17578	-0.50317	0.22073	-0.37551
12	0.17713	-0.18525	-0.12991	-0.15338	0.01760
13	-0.79267	-1.62965	0.29276	-0.63679	-2.07533
14	-0.47439	0.23371	-0.63581	-0.44355	0.0
15	-1.51222	0.0	0.35517	0.02128	0.0
16	-0.85650	-0.79777	-0.60905	0.21552	0.0
17	-0.84814	0.15454	-1.68393	-0.43280	0.0
18	0.20636	-0.65374	-0.25630	0.04556	-0.24732
19	0.10318	-0.26323	-0.93064	-1.09442	0.0
20	0.32639	-0.85410	-0.21767	-0.16122	0.0
21	-1.49371	-0.01035	0.21177	-0.36491	-0.00784
22	-0.54850	-0.53638	0.48360	-0.35363	-0.14879
23	-0.36556	0.38898	-0.68415	-0.55321	0.0
24	0.33343	-0.26922	-0.15101	-0.21802	0.0
25	-1.34718	-0.70281	-0.29600	0.05229	-2.07533
26	-0.27249	-0.81172	0.03700	-0.67307	0.0
27	-0.91479	0.06360	0.28238	-0.96467	0.0
28	-0.87096	0.11183	-0.97381	-1.28492	0.0
29	0.07043	-1.84790	-0.03267	-0.80896	0.0
30	-1.24877	-0.64532	-0.54886	0.16799	0.41554
31	0.0	-0.77073	0.15119	-0.69375	0.0
32	-0.23292	-0.07576	0.29186	-0.64619	-0.76175
33	-1.56089	-0.36582	-0.50296	0.19133	0.0
34	-0.82240	-0.88823	0.34536	-0.62651	0.0
35	-0.58083	-0.35183	0.33210	-0.93177	0.0
36	-0.62396	-1.29051	-0.16187	0.21102	0.0
37	-1.16479	0.07216	-0.49131	-0.86596	0.0
38	0.11999	-1.01578	-0.11202	-0.54876	0.0
39	0.16513	-1.29916	-0.82784	-0.42444	0.0
40	-0.72750	0.22463	-0.88104	-0.25240	-1.24977
41	0.29499	-0.64594	-0.73639	-0.92702	0.0
42	-0.11190	0.07844	-1.22892	-0.29802	0.0
43	-0.37609	-0.29875	0.22058	-0.35762	0.0
44	-0.65303	-0.75716	0.22462	-0.36885	0.0
45	0.44810	0.09913	-0.35309	-0.78659	0.0
46	0.27510	-0.50624	-1.34183	-0.75431	0.0
47	-0.46320	-0.56155	-0.16792	0.30101	0.79454
48	-0.60951	0.30546	-0.30339	-0.20025	-0.49780

Table 3
(Continued)

Item	Response Category				
	A	B	C	D	Omit
49	-2.44704	0.14546	-0.51459	-2.09941	0.0
50	0.03015	-0.45463	-0.38186	-0.94729	1.61395
51	-0.99573	0.0	-0.52498	0.06162	-1.11998
52	-1.11255	0.07756	-0.98904	-1.75495	0.0
53	-0.58026	0.05029	-0.64801	-0.82813	0.0
54	-1.51059	-1.42013	-0.42772	0.09054	-1.11998
55	0.09008	-0.71340	-1.02374	-0.07082	0.0
56	-0.34302	0.21376	-1.00782	-1.30111	1.12776
57	-0.64247	-0.57102	-0.81546	0.15782	-1.54225
58	-0.54645	0.12267	-0.27769	-0.79986	-0.89736
59	-0.34705	-0.91369	0.20322	-0.76207	1.05585
60	0.20779	-0.05873	-0.27289	-1.52205	1.61395
61	-0.44454	0.24472	-0.97139	-0.23905	0.0
62	-0.73516	-0.44950	0.09230	-0.92290	-0.49780
63	-1.25095	-1.00323	0.24507	-0.95175	1.61395
64	-0.53217	-0.12451	0.18612	-0.63249	1.61395
65	0.29456	-1.51249	-0.37689	-0.46148	0.0
66	0.35002	-0.20787	-1.17509	-0.61001	0.0
67	-0.30817	0.01781	-0.57054	0.23574	-0.14974
68	-0.48650	0.36456	-0.19889	-0.63517	0.64156
69	-0.63257	-0.53104	0.08546	0.27585	0.30675
70	-0.38289	0.28296	-0.35834	-0.05090	-0.29295
71	0.07104	-1.19163	-0.76021	0.13968	0.0
72	-0.19393	-0.23092	0.10226	0.32209	0.56966
73	-0.87246	-0.23894	-0.60794	0.33431	-1.11998
74	-1.39393	-0.82764	-0.53148	0.06830	-1.11998
75	-0.35674	-0.30673	0.53286	-0.37801	-1.25170
76	-0.87457	-0.41469	-0.65498	0.21430	-0.31066
77	-0.29718	-0.49810	0.18391	-0.40152	-0.99862
78	0.31270	-0.18129	-0.81025	-0.48366	-1.11998
79	-0.31578	-0.45368	0.36425	-0.26543	-0.44363
80	-0.84683	0.35629	-0.66554	-0.87032	0.0
81	-1.04522	-0.97891	-0.51437	0.11110	0.0
82	0.42023	-0.78210	-0.29570	-0.18339	0.03609
83	-0.16191	-0.25051	-0.79442	0.34480	-1.55094
84	-0.78919	-0.79868	0.38909	-0.92164	-0.61574
85	-0.90226	-0.96580	0.09473	-0.56422	-0.33115
86	-1.01061	-1.32039	-1.06775	0.16549	-0.33115
87	0.21552	-0.74945	-0.45423	-1.01365	-0.33115
88	0.16225	-0.54014	-0.85560	-0.51172	-0.24945
89	-1.18811	-1.74101	-0.54337	0.16161	-0.93756
90	0.00053	0.17505	-0.02515	-0.63176	0.0
91	-0.98492	0.18044	-0.95407	-0.74622	-1.31125
92	-0.60856	-0.64722	0.30119	-0.69381	-1.24377
93	-0.66083	-0.49648	-1.44547	0.42053	-0.75807
94	-0.55418	-1.16971	-0.33465	0.24625	-0.93853
95	-0.28577	-0.24132	0.30494	-0.42987	0.0
96	-0.30857	-1.37747	0.37890	-0.32830	-0.24945

Table 4

Frequency of Response to Each Response Category in Form C of the Experimental Reading Test, Sample 1R (N = 330)

Item	Response Category A	B	C	D	Omit
1	13	310	3	3	1
2	22	41	239	23	5
3	34	25	205	63	3
4	69	55	165	38	3
5	64	8	8	247	3
6	189	51	27	57	6
7	42	84	36	162	6
8	162	60	25	76	7
9	108	130	53	36	3
10	135	21	35	133	6
11	26	70	22	210	2
12	142	24	103	49	12
13	61	2	240	26	1
14	69	221	6	34	0
15	5	0	2	323	0
16	11	39	24	256	0
17	20	263	5	37	0
18	209	54	35	31	1
19	288	19	8	16	0
20	206	64	51	9	0
21	11	10	223	84	2
22	65	41	163	59	2
23	108	177	34	11	0
24	127	63	77	63	0
25	2	2	21	304	1
26	5	5	311	9	0
27	20	305	3	2	0
28	23	295	8	4	0
29	268	2	43	17	0
30	1	15	59	253	2
31	0	8	272	50	0
32	29	45	189	65	2
33	4	87	11	228	0
34	14	9	218	89	0
35	66	37	208	19	0
36	23	7	107	193	0
37	3	302	9	16	0
38	279	22	11	18	0
39	251	4	11	64	0
40	35	234	17	43	1
41	237	5	78	10	0
42	2	279	8	41	0
43	61	50	201	18	0
44	32	10	227	61	0
45	110	32	118	20	0
46	244	36	19	31	0
47	6	37	135	151	1
48	57	196	42	44	1

Table 4
(Continued)

Item	Response Category				
	A	B	C	D	Omit
49	1	270	55	4	0
50	308	16	2	3	1
51	7	0	20	302	1
52	2	312	8	8	0
53	9	305	3	13	0
54	5	2	34	288	1
55	286	10	17	17	0
56	42	250	10	17	2
57	28	20	12	268	2
58	5	268	30	25	2
59	28	9	246	46	2
60	160	76	89	4	1
61	69	200	11	41	0
62	4	19	290	16	1
63	4	19	262	44	1
64	30	67	207	25	1
65	200	7	100	23	0
66	201	25	3	101	0
67	37	30	60	192	2
68	30	159	101	39	1
69	31	54	97	145	3
70	16	132	60	110	3
71	96	6	35	193	0
72	83	35	77	83	2
73	29	45	51	204	1
74	6	10	6	307	1
75	51	34	136	106	3
76	28	36	19	245	2
77	26	44	231	27	2
78	167	101	10	51	1
79	73	26	154	69	3
80	43	228	32	27	0
81	7	9	30	284	0
82	153	42	69	62	4
83	140	46	12	131	1
84	14	74	223	16	3
85	13	10	295	11	1
86	9	16	16	288	1
87	245	17	47	18	2
88	257	38	11	22	2
89	9	9	31	279	2
90	93	105	26	16	0
91	25	275	7	21	2
92	56	19	225	29	1
93	59	43	19	207	2
94	17	16	71	223	3
95	43	39	152	76	0
96	133	8	161	21	2

Table 5

Response-Category Weights for Each of the 96 Items in Form D of the Experimental Reading Test, Sample 1R (N = 331)

Item	Response Category				
	A	B	C	D	Omit
1	-0.54981	-0.54505	-0.53226	0.12406	0.0
2	-0.50487	-0.79932	0.10705	-0.07671	0.0
3	0.38205	-0.91734	-0.79905	-0.59120	0.0
4	-0.60420	-0.29810	0.13263	-0.08021	0.63005
5	-0.43642	-0.23993	-0.39403	0.28737	-1.84406
6	-0.26631	-0.46153	0.35873	-0.21712	0.70613
7	0.07093	0.36746	-0.66703	-0.59161	0.15021
8	0.15319	0.13223	-0.95256	-0.49343	0.0
9	-0.15550	-0.28631	0.17901	0.22479	0.45666
10	0.69802	-0.14276	-0.51031	-0.48988	0.44758
11	-0.60931	0.39088	-0.48089	0.10805	0.58581
12	0.63985	-0.02133	-0.08228	-1.23783	0.0
13	-0.97354	0.05561	-1.24741	-1.13669	0.0
14	0.10088	-1.14264	-0.51748	-0.89782	0.0
15	-1.18930	0.19522	-0.49834	-0.97258	0.0
16	-0.10874	-1.09859	0.05836	0.05247	0.86597
17	0.12411	-0.20287	-0.87854	0.08205	0.0
18	-0.89185	-1.07047	-0.52499	0.14276	0.0
19	-0.38126	0.26624	-0.75293	-1.00236	-0.83634
20	0.37797	-0.04913	-0.83386	-1.02619	-0.83634
21	-0.94553	0.18567	-0.26196	-0.88976	0.0
22	0.20282	-0.36621	-0.27715	-0.17489	0.0
23	-0.58315	-0.83951	-0.59839	0.38356	0.63005
24	-0.44679	0.40406	-0.81487	-0.11845	0.0
25	0.05557	-0.75566	-0.15547	-0.91648	0.0
26	-0.59765	0.06635	-0.27488	0.0	0.0
27	0.03626	-0.04652	-0.77669	-0.73442	0.0
28	-0.96956	-0.97453	-1.01000	0.15457	0.0
29	0.28686	-0.86703	-0.75852	-0.62178	0.0
30	-0.43791	-1.15615	0.10537	-0.71526	0.0
31	-0.69006	-0.63114	-1.06000	0.24006	0.0
32	-0.72733	-0.64062	-0.52006	0.37152	0.0
33	-1.31973	-0.72043	0.21603	-0.61241	0.0
34	-0.04370	-1.00857	-0.85488	0.27028	0.0
35	-0.79579	-1.08150	0.26712	-0.77091	0.63005
36	0.25935	-1.18643	-0.30462	-0.28759	0.0
37	-1.76170	-1.42120	-0.87701	0.15016	0.0
38	0.12366	-0.57481	-1.40739	-1.56808	0.0
39	-0.39765	0.36294	-0.39453	-0.45492	0.0
40	-0.83994	0.15654	-0.36474	-1.02104	-1.58440
41	0.14206	-1.58242	-0.24172	-0.33108	0.0
42	-0.59322	-0.45172	-0.74723	0.15673	0.0
43	-0.92235	-0.71022	-0.88058	0.28471	-0.57743
44	-0.81941	-1.19724	0.24933	-0.97742	0.0
45	-0.76389	-0.96752	0.26710	-0.66772	0.0
46	-0.83582	0.33672	-0.56819	-0.69724	-0.40131
47	-0.39661	0.16604	-0.65715	-0.20865	-0.48966
48	-0.72018	-0.39770	0.35762	-0.18623	0.0

Table 5
(Continued)

Item	Response Category				
	A	B	C	D	Omit
49	-0.48946	-1.63146	-2.26497	0.04220	0.0
50	0.17035	-0.26247	-0.98577	-0.57557	0.0
51	0.07719	-0.75770	-1.43687	-1.28170	0.0
52	-1.17168	0.09267	-0.71469	-1.31195	0.0
53	-0.91422	-0.80045	0.08844	-0.72882	-0.26056
54	-0.83289	0.25225	-0.61552	-0.79147	-0.83634
55	-0.51722	-0.88269	0.17238	-1.06686	0.0
56	-1.12294	-0.95464	-1.11979	0.10706	0.0
57	-0.43003	-0.37418	0.28950	-0.30889	-1.07130
58	-1.12208	-1.46992	-1.02001	0.08512	0.0
59	-0.80843	-0.81391	-0.85472	0.29483	-2.52201
60	-0.41215	-0.77713	0.43145	-0.21132	0.0
61	-1.28154	0.23510	-0.86049	-0.73174	-0.83634
62	-0.85450	-0.98348	0.23492	-0.46082	0.0
63	-0.43123	0.23370	-1.19315	-0.87565	0.0
64	-0.64325	-0.62278	-0.63132	0.25367	0.0
65	-0.58103	-0.70261	0.26161	-0.69858	-0.91422
66	-0.55856	0.54237	-0.50306	-0.29518	0.0
67	-0.48889	0.30866	-0.52661	0.25836	0.0
68	-0.51529	-0.39812	0.11500	-0.04598	0.00123
69	-0.34100	-0.56935	-0.06931	0.46269	0.0
70	0.27758	-0.21735	-0.25805	-0.53988	0.0
71	-0.36447	0.34329	-0.55781	-0.41106	0.0
72	-0.00644	-0.45993	-1.11682	0.29089	0.0
73	-0.94453	-1.23367	0.27647	-0.64276	0.0
74	0.27428	-0.72096	-0.81910	-1.23424	-0.26056
75	-0.71445	0.22505	-0.46675	-0.67804	-0.83634
76	0.26552	-0.95000	-1.15296	-1.31763	0.0
77	0.15553	-0.55222	-0.31153	-0.38192	0.13015
78	-0.08866	0.25052	0.34916	-0.41361	0.0
79	-0.37711	0.30038	-0.75837	-0.37350	0.0
80	-0.15535	-0.36774	-0.35568	0.14306	-0.89237
81	-0.42054	-0.10806	-0.30364	0.26631	0.0
82	-0.14981	0.26748	-0.79841	-0.01587	0.0
83	-0.45690	-0.54833	0.09137	0.45889	0.0
84	0.10057	-0.77379	0.10254	-0.93839	0.0
85	-0.94678	0.12302	-0.46287	-0.97314	0.0
86	-0.95079	0.20035	-1.14898	-0.68865	0.0
87	-0.55418	0.23609	-0.38044	-0.48125	0.0
88	-1.14895	0.16977	-0.95862	-0.47943	0.0
89	-0.70460	-0.85469	0.18656	-0.65606	0.0
90	-1.21565	-1.27980	0.14934	-1.01271	0.0
91	-0.93092	-0.81923	-0.58027	0.49391	-0.17330
92	0.42951	-0.24109	-0.67909	-0.29803	0.18475
93	-0.65191	-0.94309	-0.42788	0.30133	-0.17330
94	-0.31836	-0.42421	0.40960	-0.70631	-0.26056
95	-0.84589	-0.52586	0.35162	-0.56961	-0.78341
96	-0.42227	0.17246	-0.81529	-0.23992	-1.30625

Table 6

Frequency of Response to Each Response Category in Form D of the Experimental Reading Test, Sample 1B
(N = 331)

Item	Response Category A	B	C	D	Omit
1	17	4	41	269	0
2	22	23	280	6	0
3	220	8	76	27	0
4	32	30	231	37	1
5	33	16	89	192	1
6	96	56	152	24	3
7	75	155	43	57	1
8	156	116	22	37	0
9	143	31	106	48	3
10	111	82	47	88	3
11	31	110	79	107	4
12	34	215	73	9	0
13	1	316	6	8	0
14	290	11	27	3	0
15	18	264	37	12	0
16	27	12	1	290	1
17	214	83	13	21	0
18	27	6	18	280	0
19	53	228	38	11	1
20	196	53	65	16	1
21	19	241	58	13	0
22	178	22	50	81	0
23	68	39	14	209	1
24	54	137	21	119	0
25	309	14	1	7	0
26	13	279	39	0	0
27	314	2	6	9	0
28	12	22	11	286	0
29	241	33	37	20	0
30	14	19	294	4	0
31	13	63	10	245	0
32	97	10	7	217	0
33	23	28	268	12	0
34	149	14	22	146	0
35	26	23	252	29	1
36	194	11	59	67	0
37	5	2	36	288	0
38	286	34	9	2	0
39	55	178	31	67	0
40	6	281	10	33	1
41	243	6	24	58	0
42	7	35	28	261	0
43	33	21	28	248	1
44	38	12	261	20	0
45	27	18	245	41	0
46	6	221	22	81	1
47	23	223	22	62	1
48	18	23	140	150	0

Table 6
(Continued)

Item	Response Category				
	A	B	C	D	Omit
49	3	6	1	321	0
50	245	55	23	8	0
51	307	15	5	4	0
52	8	302	15	6	0
53	1	13	296	20	0
54	21	246	35	28	1
55	33	9	269	20	0
56	8	16	7	300	0
57	25	87	188	29	2
58	8	7	8	308	0
59	18	24	42	246	1
60	62	52	174	43	0
61	14	263	33	20	1
62	18	38	258	17	0
63	61	240	11	19	0
64	29	56	10	236	0
65	10	38	240	42	1
66	56	140	23	112	0
67	57	21	56	197	0
68	23	19	203	85	1
69	168	16	3	144	0
70	176	53	63	39	0
71	69	181	25	56	0
72	147	25	24	135	0
73	19	20	251	41	0
74	248	28	46	8	1
75	8	235	60	27	1
76	262	55	7	7	0
77	232	16	49	32	2
78	46	208	15	62	0
79	44	202	32	53	0
80	38	45	20	225	3
81	31	75	68	157	0
82	30	172	50	79	0
83	105	44	31	151	0
84	228	18	68	17	0
85	9	289	10	23	0
86	37	272	9	13	0
87	37	217	63	14	0
88	13	266	11	41	0
89	34	14	263	20	0
90	7	10	292	22	0
91	21	47	65	195	3
92	161	67	61	40	2
93	42	14	56	216	3
94	47	102	167	14	1
95	12	59	206	52	2
96	67	236	8	19	1

W2 scoring procedure to the Form-D answer sheets of 328 examinees in Sample 2R-C. Similarly, the W4 weights for Form D were obtained by treating each of the 96 items in Form D as a variable for predicting total scores on Form C obtained by applying the W2 scoring procedure to the Form-C answer sheets of 331 examinees in Sample 2R-D.

Prior to calculating the partial regression coefficient for each item in Tests C and D for use in predicting total scores, the latter were converted into normalized standard scores with a mean of 50.000 and a standard deviation of 21.066. The partial regression coefficients that were obtained for scores on the 96 items in Test C are presented in Table 7. Analogous data for scores on the 96 items in Test D are shown in Table 8.

These regression coefficients based on data obtained in Samples 2R-C and 2R-D were used to modify the response-category scoring weights established for items in Test C and D on the basis of data obtained in Samples 1R-C and 1R-D. For example, the response-category W4 scoring weights for item 1 of Test C were obtained by multiplying each of the five W3 response-category weights (as shown in Table 3) by the partial regression coefficient for this item (shown in Table 7). The W4 scoring weights for the remaining 95 items in Test C and for the 96 items in Test D were obtained in an analogous manner. The resulting W4 weights are called "adjusted response-category weights." Tables 9 and 10 show the multiple correlations and associated statistical data.

Estimation of Parallel-Forms Reliability Coefficients for Total Scores on Tests C and D Obtained by Four Different Scoring Methods

To estimate parallel-forms reliability coefficients for total scores on Tests C and D that were obtained by four different scoring methods, the Form-C and Form-D answer sheets for examinees in Sample 3R (who took both forms) were used. Thus, eight total-test scores were obtained for each of the 360 examinees in Sample 3R. It should be noted that the correlation coefficients among these eight scores were based on data that had had no influence in determining the scoring weights used in methods W1, W2, W3, or W4. As a result of this cross-validation procedure, the coefficients are entirely free from spurious inflation caused by capitalization on chance effects. Mosier (1951) discussed the effects of cross-validation so they need not be presented here in detail.

Table 11 shows the four parallel-forms reliability coefficients for Tests C and D as underlined entries along with certain other intercorrelations of the eight scores obtained in Sample 3R. The underlined entries may properly be treated as reliability coefficients of either Form C or Form D because they are correlation coefficients between sets of test scores constructed to measure the same mental functions and

Table 7

Partial Regression Coefficient for Each Variable* in Form C When the Criterion Is Normalized Standard Scores on Form D
Sample 2R-C
(N = 331)

VARIABLE	B	BETA	STD ERROR B	F
VAR096	3.26072	0.06106	1.65267	3.893
VAR001	-0.33754	-0.00512	2.31162	0.023
VAR002	1.52951	0.01808	2.47210	0.383
VAR003	2.80919	0.05618	1.50142	3.501
VAR004	1.36390	0.03152	1.39812	0.952
VAR005	0.62795	0.01127	1.74398	0.130
VAR006	3.37982	0.07095	1.53518	4.847
VAR007	1.75182	0.04110	1.36926	1.637
VAR008	0.86870	0.01649	1.48791	0.341
VAR009	-3.18037	-0.06072	1.64285	3.748
VAR010	4.20629	0.09015	1.43745	8.563
VAR011	0.47842	0.00705	2.12435	0.051
VAR012	4.65864	0.03237	3.91363	1.417
VAR013	-0.84541	-0.01329	1.32879	0.405
VAR014	1.00147	0.01522	1.82534	0.301
VAR015	2.13074	0.02151	3.05651	0.486
VAR016	-1.53878	-0.02679	1.68163	0.837
VAR017	-0.69439	-0.01144	1.82953	0.144
VAR018	1.87901	0.02811	2.02534	0.851
VAR019	1.65873	0.02189	2.20537	0.566
VAR020	2.22424	0.04532	1.48159	2.254
VAR021	0.28092	0.00493	1.88268	0.022
VAR022	0.68944	0.01529	1.48169	0.217
VAR023	3.07603	0.05970	1.60991	3.651
VAR024	2.21382	0.02723	2.34312	0.893
VAR025	2.98767	0.02870	3.20719	0.868
VAR026	1.83604	0.01199	4.14349	0.196
VAR027	1.64351	0.01882	2.70878	0.368
VAR028	7.79973	0.10588	2.15242	13.131
VAR029	0.81407	0.00969	3.02204	0.073
VAR030	2.85382	0.03956	2.22149	1.650
VAR031	-1.76155	-0.02586	2.06198	0.730
VAR032	5.30592	0.09185	1.61641	10.771
VAR033	-0.05821	-0.00076	2.32066	0.001
VAR034	0.83883	0.01836	1.39880	0.360
VAR035	1.63854	0.03342	1.43649	1.301
VAR036	4.32824	0.05725	2.24292	3.724
VAR037	0.11663	0.00150	2.25083	0.003
VAR038	-2.46699	-0.03299	2.16667	1.296
VAR039	-3.12125	-0.04584	2.04644	2.325
VAR040	3.22044	0.05773	1.60530	4.025
VAR041	3.52162	0.07466	1.49866	5.522
VAR042	12.06587	0.09397	3.53824	11.629
VAR043	-1.81303	-0.02119	2.46966	0.539
VAR044	-1.56835	-0.02629	1.85887	0.712
VAR045	2.40002	0.03999	1.67927	2.043
VAR046	0.58290	0.01287	1.36299	0.183
VAR047	2.66944	0.03666	2.20269	1.469
VAR048	-2.25093	-0.03783	1.73259	1.688

* Items were considered "variables" in the multiple regression equation and are so titled.

Table 7
(Continued)

VARIABLE	B	BETA	STD ERROR B	F
VAR049	2.26965	0.03307	1.96808	1.330
VAR050	6.35137	0.06764	3.07971	4.949
VAR051	3.62703	0.03559	2.87982	1.585
VAR052	1.55958	0.02422	1.96719	0.629
VAR053	0.41689	0.00429	2.79079	0.022
VAR054	1.04462	0.01570	2.12512	0.242
VAR055	-0.06027	-0.00078	2.33450	0.001
VAR056	2.75480	0.05421	1.56022	3.118
VAR057	1.09459	0.01644	2.00075	0.299
VAR058	2.16503	0.02630	2.41725	0.802
VAR059	0.34424	0.00611	1.82491	0.036
VAR060	4.05715	0.06295	1.93457	4.398
VAR061	2.20926	0.03299	1.96543	1.264
VAR062	1.26258	0.01640	2.14386	0.347
VAR063	2.77291	0.06555	1.39393	3.957
VAR064	1.05293	0.01447	2.12074	0.247
VAR065	-0.13251	-0.00362	1.62843	0.014
VAR066	0.11603	0.00246	1.41134	0.007
VAR067	2.13551	0.03309	2.19744	0.944
VAR068	3.84105	0.06634	1.73476	4.903
VAR069	-0.92557	-0.01453	1.89569	0.238
VAR070	7.33478	0.09169	2.43288	9.089
VAR071	2.60156	0.04075	2.06345	1.590
VAR072	6.68638	0.07199	2.52200	7.029
VAR073	-0.65535	-0.01331	1.63158	0.161
VAR074	2.58859	0.03178	3.23315	0.641
VAR075	1.66588	0.03360	1.54248	1.166
VAR076	0.09027	0.00144	1.70855	0.002
VAR077	0.21778	0.00281	2.42747	0.008
VAR078	0.40037	0.00630	1.75240	0.052
VAR079	0.41177	0.00635	1.98411	0.043
VAR080	2.27293	0.05465	1.50714	2.274
VAR081	-1.54288	-0.02311	2.20864	0.438
VAR082	1.03050	0.02002	1.54705	0.444
VAR083	-1.14227	-0.01594	2.21349	0.266
VAR084	1.55909	0.04068	1.33845	1.357
VAR085	0.61803	0.00954	1.88622	0.107
VAR086	3.20104	0.06966	1.39883	5.237
VAR087	2.30051	0.03348	1.90395	1.460
VAR088	-0.29755	-0.00454	2.02402	0.022
VAR089	-0.24358	-0.00538	1.43313	0.029
VAR090	1.56724	0.02145	2.22552	0.495
VAR091	4.77590	0.08962	1.66734	8.205
VAR092	2.92034	0.05755	1.56165	3.497
VAR093	3.41470	0.08567	1.22858	7.725
VAR094	0.80070	0.01317	1.87727	0.182
VAR095	1.55703	0.02559	1.72810	0.812
(CONSTANT)	54.63094			

Table 8

Partial Regression Coefficient for Each Variable* in Form D When the Criterion Is Normalized Standard Scores on Form C
Sample 2R-D
(N = 328)

VARIABLE	B	BETA	STD ERROR B	F
VAR026	1.54291	0.03637	1.25609	0.609
VAR001	1.34301	0.01959	2.10282	0.413
VAR002	1.02732	0.01464	2.13845	0.231
VAR003	0.65830	0.01062	1.13261	0.333
VAR004	1.10973	0.01539	2.05833	0.343
VAR005	3.20524	0.06741	1.42127	5.014
VAR006	0.63133	0.01198	1.44295	0.191
VAR007	1.02008	0.03273	1.35351	0.573
VAR008	1.33746	0.02110	1.04596	0.472
VAR009	-4.55053	-0.07437	2.43445	7.243
VAR010	2.98083	0.07773	1.15965	6.607
VAR011	-1.91047	-0.03047	1.56995	1.483
VAR012	2.14989	0.03334	1.87903	1.303
VAR013	0.73023	0.01109	2.00521	0.134
VAR014	1.33818	0.02836	1.02059	0.016
VAR015	0.53453	0.01213	1.49161	0.130
VAR016	1.21045	0.01410	2.04384	0.191
VAR017	-4.33144	-0.04507	2.82184	2.240
VAR018	2.51540	0.04465	1.64302	2.341
VAR019	4.01955	0.09309	1.22963	8.130
VAR020	0.91817	0.02413	1.24062	0.548
VAR021	4.93280	0.11459	1.04350	12.427
VAR022	4.10504	0.04749	2.41070	2.900
VAR023	1.66317	0.04550	1.28625	1.673
VAR024	0.46057	0.01286	1.48859	0.107
VAR025	-0.07620	-0.00021	2.08601	0.001
VAR026	2.93215	0.02601	3.20738	0.836
VAR027	1.86624	0.01740	3.10952	0.340
VAR028	-1.59587	-0.02761	1.98377	0.647
VAR029	1.11424	0.02626	1.64716	0.458
VAR030	1.00657	0.01648	2.10222	0.272
VAR031	-0.34623	-0.01971	1.30901	0.244
VAR032	3.57960	0.09718	1.17104	9.329
VAR033	-0.30014	-0.00747	1.32817	0.051
VAR034	2.99037	0.05351	1.75209	2.913
VAR035	1.12436	0.02890	1.34046	0.684
VAR036	2.51967	0.04558	1.65565	2.314
VAR037	-0.06104	-0.00115	1.76565	0.001
VAR038	-0.82735	-0.01702	1.75238	0.223
VAR039	-0.98597	-0.02013	1.50222	0.431
VAR040	1.03349	0.02099	1.66512	0.384
VAR041	2.17997	0.03504	1.86485	1.367
VAR042	1.25204	0.02135	1.93244	0.420
VAR043	2.99155	0.07137	1.62455	3.163
VAR044	0.21509	0.00531	1.43405	0.022
VAR045	-0.25784	-0.00611	1.27623	0.035
VAR046	2.24091	0.07767	1.26325	5.415
VAR047	4.04434	0.06723	2.00132	5.500
VAR048	2.12787	0.03775	1.59633	1.777

Table 8
(Continued)

VARIABLE	B	BETA	STD ERROR B	F
VAR049	-1.21303	-0.01473	1.06532	1.264
VAR050	3.93406	0.04764	1.93020	2.398
VAR051	1.02384	0.02251	2.70034	0.503
VAR052	1.27734	0.02218	1.97620	0.443
VAR053	0.35034	0.01211	2.21364	0.148
VAR054	-3.06263	-0.06637	1.47563	4.309
VAR055	-0.60733	-0.01102	1.60182	0.144
VAR056	-1.02252	-0.02946	2.08878	0.761
VAR057	1.75044	0.03121	1.69330	1.096
VAR058	3.53001	0.04571	1.87205	1.830
VAR059	-0.70035	-0.02218	1.12855	0.502
VAR060	1.49081	0.03700	1.20870	1.521
VAR061	-4.80553	-0.12062	1.22641	15.354
VAR062	4.01801	0.00078	1.30450	0.493
VAR063	-0.11460	-0.00249	1.41545	0.007
VAR064	3.05717	0.04454	1.34883	2.325
VAR065	3.44311	0.08074	1.37883	6.235
VAR066	-2.14412	-0.05359	1.13604	3.557
VAR067	0.09063	0.00173	1.61240	0.003
VAR068	3.30777	0.04163	2.72964	2.039
VAR069	1.32303	0.02796	1.32122	1.003
VAR070	4.07601	0.07053	1.81849	7.490
VAR071	0.54563	0.01089	1.41523	0.148
VAR072	0.24151	0.00508	1.61950	0.025
VAR073	0.44453	0.01166	1.31385	0.114
VAR074	4.01033	0.08353	1.45127	7.635
VAR075	0.32947	0.00634	1.60752	0.042
VAR076	-0.67727	-0.01826	1.22000	0.304
VAR077	-4.50303	-0.05085	2.27357	4.064
VAR078	1.03226	0.01936	1.59070	0.402
VAR079	2.39498	0.04949	1.50463	2.533
VAR080	-1.14225	-0.01276	2.50933	0.207
VAR081	-0.34663	-0.01276	2.40971	0.154
VAR082	-1.23460	-0.02437	1.70344	0.010
VAR083	0.25802	0.00596	1.33250	0.037
VAR084	3.17050	0.04717	2.00553	2.283
VAR085	1.30420	0.02658	1.79610	0.603
VAR086	4.50002	0.10332	1.54744	8.836
VAR087	1.00274	0.01885	1.74434	0.332
VAR088	3.72550	0.07562	1.54303	5.830
VAR089	-0.10011	-0.00106	1.64532	0.004
VAR090	-1.75210	-0.04322	1.42403	1.503
VAR091	2.51628	0.07603	1.14025	4.794
VAR092	2.33752	0.06252	1.42215	4.342
VAR093	-1.49946	-1.01207	1.22315	0.167
VAR094	1.37543	0.02356	1.41525	1.579
VAR095	2.20213	1.05463	1.20733	2.705
(CONSTANT)	55.55525			

Table 9

Multiple Correlation and Significance-Test Summary for the Regression of Normalized Form D Standard Scores on Form C Items Sample 2R-C

Multiple R	0.93569
R Square	0.87572
Standard Error of raw-score regression plane	9.20157

Source	d.f.	Mean Square	F
Regression	30	1466.60329	[illegible]
Residual	231	84.66891	

Table 10

Multiple Correlation and Significance-Test Summary for the Regression of Normalized Form C Standard Scores on Form D Items
Sample 2K-D

Multiple R	0.93578
R Square	0.87568
Standard Error of raw-score regression plane	8.01066

Source	d.f.	Mean Square	F
Regression	96	1101.28920	17.16967
Residual	234	64.17069	

Table 11

Intercorrelations, Means, and Standard Deviations of Several Total Scores on Tests C and D Obtained by Four Scoring Methods in Sample 3R
(N = 360)

(Parallel - Forms Reliability Coefficients are Underlined)

Form	Method	C W1	C W2	C W3	C W4	D W1	D W2	D W3	D W4	Mean	SD
C	W1		.998			.882	.884			49.994	21.007
C	W2			.986	.923	.881	.883	.881	.841	49.986	21.010
C	W3				.941		.889	.894	.853	49.985	20.952
C	W4						.821	.827	.794	49.974	20.023
D	W1						1.000			49.988	20.902
D	W2							.989	.919	49.991	21.015
D	W3								.930	49.977	20.993
D	W4									49.984	20.833

expressed in normalized standard scores having similar means, standard deviations, and distributions.

It will be noted that the parallel-forms reliability coefficients increased from .882 for scoring method W1 to .883 for method W2 to .894 for method W3. It had been expected, on a a-priori grounds, that methods W1 and W2 would yield insignificantly different reliability coefficients since Tests C and D were administered under essentially untimed conditions with directions that read: "Mark items even if you are not sure of the answers, but avoid wild guessing." There was no reason to expect that variations in gambling tendencies among the examinees would markedly affect their scores.

It had, however, been expected that scoring method W3 would yield a significantly higher reliability coefficient than either of methods W1 or W2. Data from studies by Davis and Fifer (1959), Hendrickson (1971) and Reilly and Jackson (1972) supported this expectation, which was realized.

Finally, on a-priori grounds, it seemed reasonable to suppose that the "purification" of total scores likely to be brought about by scoring with method W4 would lead to obtaining a higher parallel-form reliability coefficient with scores obtained by method W4 than with scores obtained by method W3. This expectation was not confirmed by the data.

Tests of Significance of Planned Comparisons

Four planned comparisons were made to test specific hypotheses of interest. The first of these was a null hypothesis that may be written as

$$H_o: \quad \rho_{C_{W1}D_{W1}} = \rho_{C_{W2}D_{W2}} \; .$$

This hypothesis may be tested by converting both $r_{C_{W1}D_{W1}}$ and $r_{C_{W2}D_{W}}$ into their corresponding values of Fisher's z and forming a t ratio as follows:

$$t = \frac{z_{C_{W1}D_{W1}} - z_{C_{W2}D_{W2}}}{\sqrt{\left(2-2r_{z_{(C_{W1}D_{W1})}\, z_{(C_{W2}D_{W2})}}\right)\Big/ N-3)}} \; .$$

The value of the correlation coefficient between z's can be estimated in large samples by means of an equation given by McNemar (1949, p. 125, equation 48).

For the difference of .001 between the parallel-forms reliability coefficients of total scores on Tests C and D, the t ratio was .6364 with 357 degrees of freedom. Thus, the null hypothesis is accepted.

The second planned comparison was that between the reliability coefficients of Tests C and D scored by methods W2 and W3. The statistical hypotheses tested were:

$$H_o: \quad \rho_{C_{W2}D_{W2}} = \rho_{C_{W3}D_{W3}} \quad ; \text{ and}$$

$$H_1: \quad \rho_{C_{W2}D_{W2}} < \rho_{C_{W3}D_{W3}} \quad .$$

If a t ratio is formed by the same procedures used in testing the statistical significance of the first planned comparison, the value obtained is -2.3965 with 357 degrees of freedom. This result leads to rejection of the null hypothesis and acceptance (at the .01 level of significance) of the directional alternative, H_1; We conclude that response-category scoring yields a parallel-forms reliability coefficient of Tests C and D greater than does scoring with the conventional correction for chance success.

The third planned comparison was between the reliability coefficients of total scores from Tests C and D obtained by methods W2 and W4. The null hypothesis may be stated as

$$H_o: \quad \rho_{C_{W2}D_{W2}} = \rho_{C_{W4}D_{W4}} \quad .$$

If a t ratio is formed by procedures analogous to those described above, the value obtained is 6.6720. Consequently, the null hypothesis (H_o) is rejected. This leaves us in the position of concluding that total scores on Tests C and D are less reliable (at the .01 level of significance) when they are obtained by method W4 than when they are obtained by method W2. This result was not expected.

The fourth planned comparison was made between reliability coefficients for Tests C and D when scores were obtained by methods W3 and W4. The null hypothesis may be stated as

$$H_o: \quad \rho_{C_{W3}D_{W3}} = \rho_{C_{W4}D_{W4}} \quad .$$

The t ratio for testing this hypothesis is 8.3795. Again, the null hypothesis must be rejected. The data, as in the case of the third

planned comparison, ran counter to our expectations since they indicate that the reliability coefficient of Tests C and D are lower when the scores are obtained by method W4 than when they are obtained by method W3.

CHAPTER IV

THE PREDICTIVE VALIDITY STUDY

Purpose

The purpose of the predictive validity study in this investigation of the effects of differential choice weighting on test reliability and validity was to compare the predictive validity coefficients of the Davis Reading Test (Series 1, Form D) when scores were obtained by four different methods.

Test Used

The Davis Reading Test, Series 1, Form D (Davis & Davis, 1962) is designed to measure five categories of reading skills and is intended for use in grades 11 and 12 and with entering college freshmen. The test is made up of 80 items and is administered in a 40-minute time limit. Two successive equivalent scales of 40 items each are incorporated into the test. Since virtually all examinees try the first scale (40 items) in 40 minutes and very few examinees have time to try 80 items in 40 minutes, two scores can be derived from the test. The first is a Level-of-Comprehension score based on the first 40 items and the second is a Speed-of-Comprehension score based on the entire 80 items in the test. In this study only the Speed-of-Comprehension score was obtained for each examinee although the scoring weights assigned to the five choices in each item, to omissions, and to failure to reach an item in the time limit could have been used to obtain Level-of-Comprehension scores based on the first 40 items only.

Samples

As part of the regular placement testing program, Form D of Series 1 of the Davis Reading Test (Davis & Davis, 1962) was administered to freshmen upon entrance into the University of Pennsylvania. Answer sheets from this test were available for 3,840 students tested during the period 1968-1970.

Complete data, including grade-point averages at the end of their freshman year, could be located for 2,869 of the initial sample. This group, which included students from several undergraduate divisions of the University, was divided at random into three samples of 953 cases each. Random selection within undergraduate division was not done.

Thus, three groups, labeled 1V, 2V, and 3V constituted the three samples needed to conduct all steps in the investigation of the effects of weighting on predictive validity. Table 12 provides descriptive statistics pertaining to the three samples used in the predictive validity study.

Scores To Be Compared

The four methods for obtaining scores to be used in obtaining predictive validity coefficients for the Davis Reading Tests (Series 1, Form D) in Sample 3V are as follows:

W1: For each item, examinees were credited with 1 point for a correct response, 0 for an incorrect response, 0 for omission (failure to mark any choice as correct after reading the item), and 0 for not marking any choice as correct because of lack of sufficient time to consider the item. The total test score consisted of the sum of the item scores in it. This is commonly called "number-right-scoring."

W2: For each item, examinees were credited with 1 point for a correct response, $-1/(k-1)$ for an incorrect response (where k represents the number of choices per item), 0 for omission, and 0 for not marking any choice as correct because of lack of sufficient time to consider the item. This is commonly called "formula-scoring" and embodies a correction for chance success.

W3: For each item, examinees were credited with scores based on weights assigned to each choice and to the response categories of omission (failure to mark any choice as correct after reading the item) and "not read" (failure to mark any choice as correct because of lack of sufficient time to consider the item). Each scoring weight was made proportional to the mean criterion score for examinees who fell in a given response category. The criterion scores for establishing scoring weights for the Davis Reading Test were grade-point averages for examinees in Sample 1V. The total scores obtained by method W3 consisted of the algebraic sum of the scoring weights for the 80 response-categories (one per item) selected by each examinee on the Davis Reading Test.

W4: For each item, the examinees were credited with scores based on modified scoring weights assigned to each choice and to the response categories of omission and "not read." Each of the scoring weights obtained by method W3 was "modified" by multiplying it by the partial regression coefficient that would maximize the multiple correlation between a set of linear composites of the 80 item scores in the Davis Reading Test and a set of criterion scores. The criterion scores were grade-point averages for examinees in Sample 2V.

It should be noted that method W3 and method W4 differ from those described in the reliability study (p. 22). In the case of the predictive validity study the category of "not read" is considered as a valid item

Table 12

Descriptive Statistics
For Grade-Point Averages
For Three Samples of University Freshman

Descriptive Statistics	Sample 1V	2V	3V
	Raw Scores		
N	933	933	933
Mean	2.812	2.796	2.771
Variance	0.401	0.453	0.496
Standard Deviation	0.636	0.673	0.666
Range	3.670	4.000	4.000
Skewness	-0.340	-0.736	-0.621
Kurtosis	-0.197	1.252	0.397
	Standardized* Scores		
Mean	49.853	49.946	49.729
Variance	433.126	432.352	432.164
Standard Deviation	20.812	20.793	20.789
Range	111.213	103.793	106.846
Skewness	-0.041	-0.041	-0.059
Kurtosis	-0.206	-0.237	-0.212

*Note: These are normalized standard scores with mean = 50.000 and S.D. = 21.066. An approximate "table-look-up" procedure resulted in minor variations from these values for the three samples.

response category. Thus, in the present study, five choices plus omits and "not read" comprise an array of seven item-response categories for each item in the Davis Reading Test.

Determination of Scoring Weights for Method W3

Answer sheets for the 953 examinees in Sample 1V were used to obtain W3 scoring weights for the seven possible response-categories for each item of the Davis Reading Test. Criterion scores used to obtain the weights were first-semester grade-point averages after their transformation to normalized standard scores with a mean of 50.000 and a standard deviation of 21.066.

The mean criterion score of those examinees who fell in each response category for each of the 80 items in the Davis Reading Test was calculated. These means were then transformed linearly so that, within each item, the sum of the products of each transformed mean and the number of examinees entering into its calculation was made equal to zero. The transformed mean criterion score for each item-response category was used as the weight in method W3.

The W3 response-category weights for the Davis Reading Test are shown in Table 13. The numbers of examinees on which the weights are based are shown in Table 14.

Determination of Scoring Weights for Method W4

For each of the 953 examinees in Sample 2V each item of the Davis Reading Test was scored by W3 weights established in Sample 1V. With each of the 80 items considered as an independent variable in a linear composite for predicting the grade-point averages for the examinees in Sample 2V, a partial regression coefficient was obtained for scores in each predictor variable. Coefficients obtained in this manner tend to maximize the relationship between the criterion variable and the composite of variables for which the coefficients were obtained.

Prior to calculating the partial regression coefficients for each item in the Davis Reading Test for use in predicting grade-point averages, the latter were transformed into normalized standard scores with a mean of 50.000 and a standard deviation of 21.066. The partial regression coefficients that were obtained for scores on the 80 items in the Davis Reading Test are shown in Table 15.

The partial regression coefficients established on the basis of data from Sample 2V were used to modify the response-category weights established for items in the Davis Reading Test obtained using data from Sample 1V. For example, the W4 scoring weights for item 1 were obtained

Table 13

Response-Category Weights for the Davis Reading Test, Series 1, Form D
Sample 1V
(N = 953)

Item	A	B	C	D	E	Omit	NR
1	0.06216	-0.16582	-0.40902	-0.15656	-0.02228	-0.29207	0.0
2	-0.28671	-0.43235	-0.10778	0.10972	-0.50426	-0.19942	0.0
3	-0.15133	0.03423	-0.16504	-0.31066	-0.24721	-0.19994	0.0
4	0.03352	-0.16317	0.05502	-0.16779	-0.15075	-0.34685	0.0
5	-0.00409	0.05862	-0.08962	-0.17371	-0.22772	-0.57473	0.0
6	-0.36670	0.00388	-0.17564	0.04129	-0.08176	0.08252	0.0
7	-0.22824	0.00871	0.05879	-0.07430	-0.09079	-0.47273	0.0
8	-0.07761	-0.07564	-0.06063	0.02361	-0.63716	0.05737	0.0
9	-0.27277	0.23692	-0.05835	0.07594	0.12365	-0.35178	0.0
10	-0.18502	0.04252	-0.40199	0.01945	0.04740	-0.03997	0.0
11	0.02841	-0.21331	-1.09550	-0.19584	-0.03904	-0.19243	0.0
12	-0.07842	-0.27269	0.01630	-0.37126	0.06430	-0.06862	0.0
13	-0.16859	-0.33682	-0.25798	-0.34443	0.01328	0.07626	0.0
14	0.05155	-0.06952	-0.30601	-0.16373	-0.15953	-0.08716	0.0
15	-0.26549	-0.25978	0.04125	0.05532	-0.11091	-0.13851	0.0
16	-0.09663	-0.79716	-0.14934	-0.64233	0.08641	-0.16953	0.0
17	0.02959	-0.14466	-0.07975	-0.46625	0.17042	0.41596	0.0
18	-0.33614	0.01282	-0.38007	-0.04143	0.12854	0.0	0.0
19	-0.02028	-0.07577	0.02361	-0.31096	-0.07420	-0.33472	0.0
20	-0.50815	-0.09722	-0.06692	0.36648	0.03411	-0.83621	0.0
21	-0.03275	-0.15159	-0.04203	0.05045	-0.32259	-0.63038	0.0
22	-0.11405	0.01550	-0.05241	-0.20499	0.03537	-0.19221	0.0
23	0.07007	-0.05006	0.02156	-0.22051	-0.23692	-0.16469	0.0
24	-0.15080	-0.15697	-0.67988	-0.19304	0.04544	0.05012	0.0
25	-0.07462	-0.20469	-0.02571	0.05853	-0.03749	0.08673	0.0
26	-0.22784	-0.20030	0.04413	-0.11149	0.00322	-0.35383	0.0
27	-0.18247	-0.24577	0.04506	-0.01843	0.09591	-0.24257	0.0
28	-0.53969	0.08376	-0.11316	-0.08742	0.59107	-0.20815	0.0
29	-0.14190	0.23242	-0.33294	0.03830	-0.10125	-0.33694	0.0
30	0.07720	-0.33258	-0.21669	-0.09926	-0.47953	-0.29080	0.0
31	-0.03955	0.04271	-0.25676	-0.24239	0.01293	-0.15544	0.0
32	-0.10923	-0.38765	0.04701	-0.45956	0.0	-0.27311	0.0
33	-0.13829	-0.14259	-0.35610	0.17128	-0.13829	-0.24335	0.0
34	-0.15645	-0.06006	-0.27765	-0.02981	0.12132	-0.33039	0.0
35	-0.03860	-0.25804	0.09577	0.07729	-0.00615	-0.04665	0.0
36	-0.03813	0.05758	-0.04959	-0.28438	-0.12081	-0.13978	0.0
37	-0.01297	-0.40360	0.02625	-0.25930	0.03644	-0.02359	0.0
38	0.03873	-0.09018	0.14533	-0.33885	0.03629	-0.06647	0.0
39	0.07111	-0.13913	-0.25266	-0.33848	0.07808	0.07792	0.0
40	-0.32077	0.01343	-0.01027	0.0	0.06048	-0.14217	0.0

Table 13
(Continued)

Item	A	B	C	D	E	Omit	NR
41	-0.67891	-0.01984	0.09603	-0.21993	0.05203	-0.16732	0.0
42	-0.35374	0.03594	0.04779	-0.15854	-0.41790	0.04605	0.24142
43	-0.40755	-0.33954	-0.14960	-0.37820	0.05899	-0.26554	-0.97367
44	0.06909	-0.32986	-0.40011	-0.12630	-0.09597	-0.40257	-0.97367
45	-0.12963	-0.35366	-0.23187	0.07363	-0.47207	-0.35484	-0.97599
46	0.10283	0.10715	0.57011	0.00968	-0.44167	-0.47705	-1.06974
47	0.00894	-0.44635	-0.23256	0.05247	0.03669	-0.31535	-1.06974
48	-0.55733	0.23572	0.05530	-0.02643	-0.07012	-0.16238	-0.93070
49	-0.36724	0.04848	-0.04958	0.08127	-0.45232	-0.11243	-0.80167
50	0.89905	-0.04727	0.06456	0.10458	-0.20873	0.01951	-0.80167
51	-0.13557	-0.06935	-0.12149	-0.12877	0.07935	0.06568	-0.64539
52	-0.24713	-0.09022	0.09803	-0.23560	0.02365	-0.06777	-0.54791
53	0.03197	-0.04868	0.00151	0.09765	0.02904	0.00654	-0.57020
54	-0.07480	0.12721	-0.08705	-0.16914	-0.07809	-0.12193	-0.63522
55	-0.41011	0.22881	-0.26511	-0.18194	0.05664	-0.06473	-0.64869
56	0.10461	-0.02480	-0.32265	0.09561	0.02811	-0.08699	-0.35905
57	-0.05630	0.11855	-0.48907	-0.31942	-0.02541	-0.23378	-0.34266
58	-0.45695	0.01428	-0.14377	0.08326	-0.17633	-0.05098	-0.35172
59	-0.27331	0.09119	-0.29762	0.23405	-0.11021	-0.03752	-0.35069
60	-0.04611	-0.16254	-0.22453	-0.18458	0.10536	-0.16255	-0.33700
61	-0.19810	-0.28361	-0.18273	0.15300	-0.13296	-0.12647	-0.30264
62	0.08799	0.20013	-0.07378	-0.19609	-0.05901	-0.16969	-0.27455
63	-0.01818	-0.17005	0.15878	0.09305	0.11011	-0.16430	-0.18170
64	0.13768	-0.12939	-0.01250	-0.27769	-0.15023	-0.11638	-0.13076
65	0.06893	0.41735	0.10682	-0.08071	-0.21904	-0.18979	-0.08507
66	-0.38439	-0.04401	0.21196	0.05575	-0.50749	-0.08041	-0.07211
67	-0.20562	-0.19213	-0.10414	0.26940	-0.11593	-0.09323	-0.05805
68	0.12448	0.00797	-0.00606	-0.44393	0.13683	-0.18106	-0.06539
69	-0.06813	-0.35517	0.10416	-0.04899	-0.09323	-0.03659	-0.06839
70	-0.25335	0.05313	-0.12748	0.12612	-0.21775	-0.08635	-0.05186
71	0.12854	-0.42784	0.17785	0.35964	0.12963	-0.22493	-0.07086
72	-0.11744	0.13562	0.51396	-0.35786	0.01536	-0.09967	-0.06885
73	-0.03536	0.24522	-0.21827	-0.06479	-0.25279	-0.04401	-0.07254
74	0.17820	-0.29786	-0.05584	-0.10436	0.14664	-0.08625	-0.07965
75	0.24142	-0.52249	0.15486	0.15599	-0.17688	-0.43231	-0.06934
76	0.04429	0.12718	0.61886	-0.57535	-0.01871	0.06805	-0.02117
77	-0.09164	0.17524	0.16626	-0.23399	-0.93584	0.14269	-0.03105
78	0.01473	-0.40787	0.16128	0.02709	0.50635	-0.23840	-0.02693
79	-0.40711	0.04193	0.12016	0.18016	0.06722	-0.38095	-0.01570
80	0.16715	-0.10316	0.0	-0.00178	0.12207	0.0	-0.01448

Table 14

Frequency of Response to Each Response Category
in the Davis Reading Test, Series 1, Form D
Sample 1V
(N = 953)

Item	A	B	C	D	E	Omit	NR
1	574	22	9	110	217	21	0
2	54	18	306	561	8	6	0
3	101	803	8	12	27	2	0
4	794	47	3	26	72	11	0
5	211	530	142	50	8	12	0
6	20	87	62	634	128	22	0
7	21	877	32	7	9	7	0
8	13	44	91	786	14	5	0
9	142	24	93	612	44	38	0
10	109	773	33	4	22	12	0
11	807	29	2	49	51	15	0
12	96	31	22	79	713	12	0
13	25	16	10	3	877	22	0
14	617	254	16	5	42	19	0
15	37	23	509	245	53	81	0
16	193	5	56	16	666	17	0
17	782	88	54	18	9	2	0
18	12	871	15	51	4	0	0
19	24	48	789	20	61	11	0
20	4	104	123	11	702	9	0
21	43	21	236	609	28	11	0
22	44	274	71	46	499	19	0
23	105	136	643	39	11	19	0
24	92	15	14	110	717	5	0
25	142	28	75	457	225	26	0
26	19	23	725	94	60	32	0
27	192	45	145	21	528	22	0
28	9	527	291	86	12	28	0
29	76	4	16	779	48	28	0
30	717	44	29	72	4	87	0
31	20	765	73	31	27	37	0
32	55	39	802	5	0	52	0
33	41	335	38	475	12	52	0
34	17	57	11	508	318	42	0
35	162	64	527	52	100	48	0
36	26	532	297	6	68	24	0
37	9	12	814	66	34	18	0
38	627	59	54	79	87	47	0
39	632	232	49	8	21	11	0
40	23	854	11	0	49	16	0

Table 14
(Continued)

Item	A	B	C	D	E	Omit	NR
41	13	46	32	124	718	20	0
42	15	26	761	98	41	11	1
43	32	38	9	38	817	17	2
44	730	70	6	24	94	27	2
45	12	1	116	764	36	21	3
46	25	81	5	801	21	15	5
47	25	3	49	4	827	40	5
48	21	8	638	89	147	44	6
49	33	810	51	8	39	3	9
50	1	296	558	5	78	6	9
51	129	36	116	25	567	64	16
52	20	88	588	131	42	65	19
53	7	39	559	88	121	118	21
54	45	507	126	112	24	115	24
55	23	37	52	17	717	80	27
56	252	162	8	58	247	180	46
57	255	499	7	10	56	61	65
58	15	39	52	658	65	47	77
59	22	666	78	21	18	56	92
60	40	22	4	103	633	55	96
61	31	11	40	530	111	109	115
62	55	412	86	69	26	171	134
63	42	82	232	277	27	140	153
64	458	154	16	12	21	126	166
65	578	2	45	4	80	65	179
66	11	58	24	511	2	82	265
67	83	59	36	265	77	145	288
68	270	111	129	13	47	80	303
69	41	31	409	29	10	102	331
70	22	6	46	373	34	115	357
71	305	17	33	5	87	97	409
72	16	209	19	19	105	96	429
73	86	238	33	61	38	55	442
74	8	15	53	23	340	48	461
75	1	2	366	3	71	25	487
76	229	63	4	3	85	5	564
77	142	131	17	9	2	7	595
78	50	13	62	177	19	13	619
79	5	31	36	12	193	22	654
80	0	8	0	192	81	0	666

Table 15

Partial Regression Coefficients for Scores of 80 Items
in the Davis Reading Test, Series 1, Form D,
for Predicting Freshman Grade-Point Averages
Sample 2V
(N = 953)

VARIABLE	B	BETA	STD ERROR B	F
VAR000	0.61413	0.03628	0.60125	1.043
VAR001	-0.45830	-0.05987	0.23584	3.454
VAR002	-0.03856	-0.02202	0.15242	0.418
VAR003	-0.30143	-0.03485	0.32809	0.844
VAR004	-0.15883	-0.01852	0.28869	0.305
VAR005	0.13712	0.02036	0.23307	0.346
VAR006	0.38427	0.05108	0.27718	1.922
VAR007	-0.72027	-0.05406	0.43656	2.722
VAR008	-0.00934	-0.00089	0.37235	0.001
VAR009	0.16341	0.03470	0.15419	1.123
VAR010	0.12537	0.02096	0.20178	0.374
VAR011	0.14573	0.02010	0.25471	0.327
VAR012	0.11999	0.02435	0.16259	0.545
VAR013	0.12444	0.01226	0.36300	0.118
VAR014	-1.37500	-0.14824	0.33877	16.474
VAR015	0.35815	0.05075	0.21884	2.368
VAR016	0.09597	0.02200	0.14859	0.417
VAR017	0.05858	0.00825	0.24536	0.057
VAR018	0.52431	0.04393	0.38097	1.894
VAR019	0.07763	0.00759	0.35986	0.047
VAR020	-0.06100	-0.01051	0.19592	0.097
VAR021	0.11179	0.01851	0.19895	0.316
VAR022	0.22649	0.02317	0.32267	0.493
VAR023	0.00611	0.00057	0.39234	0.000
VAR024	0.05549	0.00918	0.21420	0.067
VAR025	0.20465	0.02143	0.31834	0.413
VAR026	-0.10349	-0.01428	0.25627	0.163
VAR027	0.11124	0.02066	0.18975	0.344
VAR028	0.04780	0.00928	0.17949	0.071
VAR029	0.01649	0.00217	0.26752	0.004
VAR030	0.50009	0.05907	0.18189	2.722
VAR031	0.38129	0.05228	0.26255	2.109
VAR032	0.04663	0.00817	0.21101	0.049
VAR033	0.29914	0.07911	0.12860	5.411
VAR034	-0.19346	-0.03004	0.21618	0.801
VAR035	0.36551	0.04220	0.27662	1.746
VAR036	0.49401	0.05165	0.34322	2.072
VAR037	0.01961	0.00260	0.27374	0.005
VAR038	0.46988	0.08391	0.18604	6.379
VAR039	0.41329	0.06952	0.21326	3.756
VAR040	-0.07537	-0.01005	0.24341	0.096

Table 15
(Continued)

VARIABLE	B	BETA	STD ERROR B	F
VAR041	0.32761	0.05586	0.18967	2.984
VAR042	-0.03027	-0.00535	0.19543	0.024
VAR043	0.07685	0.01564	0.16738	0.212
VAR044	0.00168	0.00040	0.17149	0.000
VAR045	-0.07933	-0.01788	0.16772	0.224
VAR046	-0.37191	-0.07431	0.20994	3.138
VAR047	-0.06779	-0.01255	0.23118	0.086
VAR048	0.15068	0.02516	0.23149	0.346
VAR049	0.42801	0.09012	0.17884	5.728
VAR050	0.55190	0.10533	0.21408	6.646
VAR051	0.18684	0.03492	0.22091	0.715
VAR052	0.16704	0.03581	0.17191	0.944
VAR053	-0.13564	-0.02500	0.36990	0.252
VAR054	-0.05775	-0.01352	0.18750	0.095
VAR055	0.15660	0.03607	0.18613	0.708
VAR056	0.03381	0.00556	0.26315	0.017
VAR057	0.32452	0.07239	0.17701	3.361
VAR058	0.21736	0.05089	0.17115	1.613
VAR059	-0.35038	-0.08320	0.20014	3.074
VAR060	-0.55998	-0.13176	0.20669	7.340
VAR061	0.34976	0.09220	0.15852	4.868
VAR062	0.14410	0.04090	0.15209	0.898
VAR063	-0.04645	-0.00990	0.19824	0.055
VAR064	0.09836	0.01944	0.19654	0.250
VAR065	-0.26262	-0.04422	0.22091	1.413
VAR066	-0.29627	-0.03880	0.28767	1.061
VAR067	0.21959	0.05657	0.14529	2.285
VAR068	0.32707	0.05243	0.25081	1.701
VAR069	-0.48299	-0.07488	0.25900	3.478
VAR070	-0.14568	-0.02304	0.27059	0.290
VAR071	0.04720	0.01013	0.18769	0.065
VAR072	0.39788	0.07630	0.20900	3.624
VAR073	0.18681	0.04182	0.17443	1.147
VAR074	-0.20862	-0.03603	0.27243	0.586
VAR075	0.00724	0.00149	0.20245	0.001
VAR076	0.24261	0.02162	0.41599	0.340
VAR077	-0.23319	-0.03677	0.22766	1.049
VAR078	0.21743	0.03058	0.24271	0.803
VAR079	0.57212	0.06510	0.29006	3.890
(CONSTANT)	2.78251			

by multiplying each of the seven W3 response-category weights (as shown in Table 13) by the raw-score partial regression coefficient for item 1 (shown in Table 15). The W4 weights for the remaining items were obtained in the same way. The resulting W4 weights are termed "adjusted response-category weights." Table 16 shows the multiple correlation and associated statistical data.

Estimation of Predictive Validity Coefficients for the Davis Reading Test Total Scores Obtained by Four Different Scoring Methods

To estimate predictive validity coefficients for total score on the Davis Reading Test obtained by four different scoring methods, the answer sheets for examinees in Sample 3V were used. Four total-test scores were obtained for each of the 953 examinees in Sample 3V. Sample 3V in no way influenced the determination of either response-category weights or partial regression coefficients. Thus, the correlation coefficients are free from spurious inflation caused by capitalization on chance effects.

The predictive validity coefficients, as shown in the first row of Table 17, were obtained by correlating total-test scores of the Davis Reading Test by four different scoring methods with grade-point averages for the examinees in Sample 3V. The grade-point averages had been transformed into normalized standard scores with a mean of 50.000 and a standard deviation of 21.066.

The predictive validity coefficients appear to be quite similar for methods W1, W2, and W3. It had been expected, however, on a-priori grounds, that method W2 would yield a higher validity coefficient than W1.

It had also been expected that method W3 would result in a predictive validity coefficient higher than W2. The intent of previous studies (Davis & Fifer, 1959; Hendrickson, 1971; Reilly & Jackson, 1972) was to improve reliability through techniques of response-category weighting similar to those employed here. None of these studies sought to improve predictive validity directly, however. It would seem, though, that the same line of reasoning would apply. The weighting procedure, as defined here, tends to maximize the relationship between item scores and a criterion. If the criterion of interest is grade-point averages, weighting a test to predict that criterion should tend to maximize predictive validity.

It seemed reasonable to hypothesize that, if the W3 scoring method tended to maximize the relationship between the criterion and a test weighted by that method, the coefficient for W3 would be greater than for, say, W1 or W2. This expectation was not realized.

Table 16

Multiple R for Regression of GPA on the
80-item Davis Reading Test, Series 1, Form D
Sample 2V

Multiple R	0.42027
R Squared	0.17663
Standard error of raw-score regression plane	0.64277

Source	d.f.	Mean Square	F
Regression	80	0.96604	2.33824
Residual	372	0.41315	

Table 17

Product-Moment Intercorrelations Among Grade-Point Averages and Davis Reading Test Scores Obtained by Four Scoring Methods in Sample 3V
(N=953)

Variable	GPA	W1	W2	W3	W4
GPA	1.000	.297	.332	.298	.407
W1		1.000	.897	.876	.429
W2			1.000	.540	.437
W3				1.000	.335
W4					1.000
Mean*	49.946	49.998	50.000	49.989	49.997
SD*	20.792	20.982	20.982	20.994	21.012

*Note: Scores on all variables are expressed as standiles (Mean = 50.000, S.D. = 21.066). Due to a "table-look-up" procedure, minor variations occurred in this transformation.

Finally, it seemed reasonable to suppose that W4 scoring would lead to a higher reliability coefficient then W2 scoring or W3 scoring. These expectations were realized.

Tests of Significance of Planned Comparisons

Four planned comparisons were made. Each comparison tested a specific hypothesis of interest. The first hypothesis was one of no difference between predictive validity coefficients when the Davis Reading Test was scored by methods W1 and W2. That is, H_o: $\rho_{(GPA)(W1)} = \rho_{(GPA)(W2)}$. The statistical significance of the difference between the two coefficients can be obtained by applying the equation

$$t = (r_{12}-r_{13}) \sqrt{\frac{(N-3)\ (1+r_{23})}{2(1-r_{12}^2-r_{13}^2-r_{23}^2+2r_{12}r_{13}r_{23})}} \ .$$

This equation (McNemar, 1949) takes into the consideration the fact that the correlations being compared were obtained in the same sample and are themselves correlated or dependent.

For the first planned comparison a t value of less than unity was obtained, which indicates that the difference is not statistically significant. The hypothesis of no difference between the two correlation coefficients could not be rejected.

The second planned comparison tested a hypothesis of no difference between predictive validity coefficients obtained when the Davis Reading Test was scored by method W2 and by method W3. The hypothesis, stated in null form, was: H_o: $\rho_{(GPA)(W2)} = \rho_{(GPA)(W3)}$. The appropriate equation was applied. The t value was found to be less than unity. Again, the hypothesis of no difference could not be rejected.

The third planned comparison tested the hypothesis of no difference between $\rho_{(GPA)(W2)}$ and $\rho_{(GPA)(W4)}$. The comparison resulted in a t value of 3.397. The probability that a difference in correlation coefficients as great as that obtained would occur by chance is less than .05. We conclude that the predictive validity of the Davis Reading Test was significantly improved by using W4 scoring instead of the conventional W2 method.

The fourth planned comparison was made between predictive validity coefficients when the Davis Reading Test was scored by methods W3 and W4. The null hypothesis is stated as:

$$H_o: \ \rho_{(GPA)(W3)} = \rho_{(GPA)(W4)} \ .$$

The obtained t ratio of 3.858 is statistically significant (d.f. = 950; $p < .05$). The null hypothesis must be rejected, and we conclude that an improvement in the predictive validity of the Davis Reading Test for freshmen first-semester grade-point averages can be obtained by the use of the modified response-category weights yielded by method W4. While W3 scoring does not lead to increases in predictive validity, the W4 method does. W4 scoring involves both Guttman response-category weighting and item weighting (based upon multiple linear regression procedures) and alters the characteristics of the scores in such a way as to maximize their predictive validity for a designated criterion variable.

CHAPTER V

SUMMARY, DISCUSSION, AND CONCLUSIONS

Summary of the Reliability Study

Investigation of the effects of various weighting methods on test reliability and predictive validity are reported in the literature periodically. Several recent studies (Davis & Fifer, 1959; Hendrickson, 1971; Reilly & Jackson, 1972) have reported mixed results using a differential choice weighting procedure similar to the one used in this investigation.

The purpose of the reliability study was to compare the parallel-forms reliability coefficients of two forms (C and D) of an experimental reading-skills test when scores were obtained by four different methods. Two of the methods were: 1) "number-right scoring" where, for each item, examinees received 1 point for a correct response and 0 for any other response, and; 2) "formula-scoring" that involved a correction for chance success. For each item, examinees received 1 point for a correct response, 0 for omission, and $-1/(k-1)$ points for an incorrect response. These scoring methods do, in a sense, weight the response alternatives differentially. Both are commonly employed in the scoring of aptitude tests and require no explanation of the background upon which they are based. In the reliability study these test-scoring methods were labeled W1 and W2, respectively.

Two other methods of test scoring under study in this investigation involved the differential weighting of item choices. The two methods were: 1) response-category-weight scoring which involved cross-validated weights for every item response category including omission, and; 2) "adjusted" response-category-weight scoring which involved cross-validated weights for every item-response category including omission after the weights have been adjusted by means of cross-validated partial regression coefficients for predicting a defined criterion. These test-scoring methods were lableled W3 and W4, respectively.

The method of weighting item-response categories that was used in this investigation was described some time ago by Guttman (1941). Guttman showed that to maximize the relationship between a criterion and the response categories for any given test item the weight for each category should be linearly related to the mean criterion score of persons who select that category. This weighting procedure was one of the test-scoring methods used in this study (labeled W3) and was the basis for another (labeled W4).

Although the procedure described by Guttman leads to a relationship

between a criterion and the response categories in a single test item that is at a maximum, it does not necessarily lead to a relationship that is maximized for that criterion and a series of test items. When scores for a series of test items are to be summed to obtain a total-test score for a person, the relationship between the total-test scores and criterion scores will tend to be at a maximum when each item is weighted by the appropriate partial regression coefficient. The W4 test-scoring method consisted of the procedure described by Guttman plus the multiple regression procedure. The combination of the two weighting procedures leads to response-category weights for an item that are "adjusted" by the partial regression coefficient for the item containing the response categories.

Methods W3 and W4 required that, for each examinee, a total-test score on both forms of the test be available. This was necessary because the weight determined for each response category of an item in a test form was proportional to the mean score on the corresponding parallel form of all examinees selecting that response category, Thus, the mean total-test score on Form D of those examinees who fell in each response category for each item in Form C of the test was calculated. Analogous scoring weights were obtained for the categories in Form D.

In another sample the test items in each form for each examinee were scored using the obtained response-category weights. Each of the 96 items in one form, scored by response-category weights, was treated as an independent variable for predicting total score on the corresponding parallel form of the test. The regression of the parallel-form test score upon the 96 items in the corresponding form produced a partial regression coefficient for each item in Tests C and D. The weight for each response category within a test item was multiplied by the partial regression coefficient for the item. The products were termed "adjusted response-category weights." This procedure provided the weights required for the W4 method of scoring.

In another sample of examinees for whom data on both Forms C and D were available, scores for each test form for each examinee were obtained by methods W1, W2, W3, and W4. Certain intercorrelations of the eight scores may be interpreted as parallel-forms reliability coefficients for Forms C and D scored by each of the four methods.

Statistical comparisons revealed that for Form C and D:

1) No difference was found between the parallel-forms reliability coefficients of "number-rights" scores (method W1) and scores corrected for chance success (method W2);

2) Scoring with W3 weights for each response category resulted in a significant increase in parallel-forms reliability over that of scoring with a correction for chance success (method W2);

3) Scoring with W4 weights for each item choice yielded a reliability coefficient for the resulting scores that was significantly lower than the reliability coefficient for scores corrected for chance success (method W2);

4) Scoring with W4 weights for each item choice yielded scores significantly less reliable than scores yielded by method W3.

Summary of the Predictive Validity Study

The objective of the predictive validity study was to compare the predictive validity coefficients of the Davis Reading Test (Series 1, Form D) when scores were obtained by four different test scoring methods. The criterion scores in this study were first-semester grade-point averages for university freshmen. The four test-scoring methods compared were: 1) number-right scoring; 2) scoring using a correction for chance success. These two scoring methods are identical to those used in the reliability study; 3) scoring with weights for each response category plus omission and "not read" (omitting an item because of lack of sufficient time to consider the item), and 4) scoring with weights for every response category for each item "adjusted" by the appropriate partial regression coefficient. These methods were labeled W1, W2, W3, and W4, respectively.

The predictive validity study differed from the reliability study in two important ways. First, the response-category weighting procedures differentiated between omission and "not read" (failure to mark any choice as correct because of lack of sufficient time to consider the item). Second, the criterion scores used to determine response-category weights were not test scores, but were first-semester grade-point averages for freshmen at the University of Pennsylvania.

Three samples of examinees were required in the predictive validity study. Using method W3, response-category weights were established in one sample according to the Guttman response-category weighting procedure described elsewhere in this report. In a second sample drawn from the same parent population, grade-point averages were regressed, in a multiple linear regression, on scores for each of the 80 items in the Davis Reading Test. Each of the test items had been scored using the response-category weights obtained in the first sample of examinees. The required "adjusted response-category weights" were obtained by multiplying each weight in an item by the partial regression coefficient for that item. This procedure had been labeled method W4. Since each step in the determination of the response-category weights and regression coefficients involved an independent sample of examinees, the obtained weights and coefficients were free from spuriousness caused by capitalization on errors within the samples in which the weights and coefficients were obtained. Scores on the Davis Reading Test in a third sample of examinees were obtained using the four test-scoring methods. Predictive validity coefficients for the test scored in each manner were obtained by correlating test scores with grade-point averages.

Planned statistical comparisons between selected pairs of validity coefficients revealed that:

1) No significant difference was found in the predictive validity for "number-right" scores (method W1) and scores corrected for chance success (method W2);

2) No significant difference was found in the predictive validity of scores obtained by applying W3 weights for each item-response category and for scores corrected for chance success (method W2);

3) Scoring with "adjusted response-category" weights (method W4) resulted in a significantly higher predictive validity coefficient than scoring with a correction for chance success (method W2);

4) Scoring with "adjusted response-category" weights (method W4) resulted in a significantly higher predictive validity coefficient than scoring with W3 weights for each response category.

Discussion and Conclusions of the Reliability Study

As shown in Table 11, the parallel-forms reliability coefficients of scores obtained by scoring methods W1, W2, W3, and W4 were .882, .883, .894, and .794, respectively.

The fact that methods W1 and W2 yielded scores that were virtually identical with respect to their reliability coefficients had been expected because the tests had been administered under generous time limits that permitted every examinee to consider every item and because the directions included the sentence "Mark items even if you are not sure of the answers, but avoid wild guessing."

Because the use of differential choice weights obtained by Guttman's procedure (Guttman, 1941) allows the variance generated by use of partial information and misinformation in the marking of answers to items to which an examinee is not sure of the correct answer to be included in test scores, it was expected that the reliability coefficient of W3 scores would be higher than that of either W1 or W2 scores. This expectation was realized.

On the other hand, the a-priori expectation that W4 scores would have a higher reliability coefficient than W3 scores was not realized. Instead, as noted above, the parallel-forms reliability coefficient of W4 scores was significantly lower than that of the W3 scores by about .1. An adequate explanation of this phenomenon has simply not as yet been formulated.

In general, it may be concluded that differential choice weights for item-response categories are useful for improving reliability per unit of test length. This confirms most previous studies pertaining to this point (Davis & Fifer, 1959; Hendrickson, 1971; Reilly & Jackson, 1972).

Discussion and Conclusions of the Predictive Validity Study

Table 17 shows that the validity coefficients of scores obtained by methods W1, W2, W3, and W4 are .297, .302, .298, and .407 respectively.

The a-priori expectation that W2 scores would be more valid than W1 scores was not realized. Changes in predictive validity produced by scoring with a correction for chance success are usually small and, unless very large numbers are available, it is difficult to demonstrate statistically significant differences. Although the difference in the validity coefficients as a result of W2 versus W1 scoring was positive ($\sim$.003), the test was not statistically significant. The lack of a significant difference is due, in part, to the high correlation between the two types of scores. Large differences must occur for the difference to be statistically significant.

The directions for the Davis Reading Test include a statement against guessing wildly from among the choices if the correct answer is not known. Because of this the behavior of some examinees tends to be more cautious thus eliminating some variance in the scores due to guessing. This effect would apply to "number-right" scores (W1) as well as the "formula score" (W2).

The expectation that W3 scores would be more valid than W2 scores was not realized. One reason for the lack of improvement in the predictive validity as a result of W3 scoring might well be due to the importance that omitted and "not read" items assume in the weighting scheme. The Speed score in the Davis Reading Test indicates basically the rapidity and accuracy with which the examinee understands the kinds of material ordinarily required at the college level. Perhaps the W3 method alters the characteristics of the test in such a way as to increase the importance of the speed factor. W3 scoring might "refine" the measurement of speed of comprehension to a much greater extent than that obtained by either W2 or W1 scoring. Hendrickson (1971) has suggested that the factor structure of a test might be altered as a result of Guttman response-category weighting. Further, speed of comprehension as a reading skill may account for less variance in the criterion (grade-point average) than other factors measured in the weighted test.

The decrease in predictive validity that seemed to result from Guttman response-category weighting using grade-point average as the criterion was compensated for by W4 scoring. By "adjusting" the response-category weights by the appropriate partial regression coefficients, the effects balance each other out. W4 scoring weights more heavily those items that account for the greatest amount of variance in the criterion. Improvement in test validity due to W4 scoring was expected.

The use of response-category weighting rests upon the important

consideration that the item options be sufficiently well-written and refined to accurately measure the various degrees of partial information held by examinees. Davis (1959) has emphasized the point that improvement in reliability and, presumably validity, is attributed to the selection among incorrect options by examinees who are unable to select the keyed option.

With regard to response-category weighting and item weighting several points must be considered. First, weights should be established using large samples of examinees to insure stability of the weights upon cross-validation. Second, consideration should be given to the magnitude of the weights assigned to incorrect and omit categories. The dilemma posed when the weight for the keyed category is less than the weight for an incorrect category should be resolved. This point is especially important in light of the comments by Green (1972) about the ethical problems posed by directions about omitting items when in doubt. Frederick B. Davis (personal communication) has suggested that the test directions should convey to examinees the nature of the test scoring procedure. Davis has also suggested that the correct category, omit and, perhaps "not read" categories each receive standard weights and incorrect categories receive differential weights. A refinement of the empirical weights through a procedure similar to this might overcome the ethical problems cited by Green (1972).

Although the results of the reliability and predictive validity studies are mixed, the evidence points to the value of response-category weighting for improving test reliability. The value of response-category weighting for improving predictive validity is less apparent. The application of response-category weighting with item weighting holds promise as a means of improving predictive validity. Further research in this area is, however, required before a definitive statement about the overall value of response-category weighting can be made.

REFERENCES

Coombs, C. H., Milholland, J. E., & Womer, F. B. The assessment of partial knowledge. Educational and Psychological Measurement, 1956, 16, 13-37.

Corey, S. M. The effect of weighting exercises in a new-type examination. Journal of Educational Psychology, 1930, 21, 383-385.

Davis, F. B. Estimation and use of scoring weights for each choice in multiple choice test items. Educational and Psychological Measurement, 1959, 19, 291-298.

Davis, F. B. Analyse des items. Louvain, Belgium: Nauwelaerts, 1966.

Davis, F. B. Research in comprehension in reading. Reading Research Quarterly, 1968, 3, 499-545.

Davis, F. B., & Davis, C. C. The Davis reading tests, Series 1 and 2, Forms A, B, C, D. New York: Psychological Corporation, 1962.

Davis, F. B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. Educational and Psychological Measurement, 1959, 19, 159-170.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 1965, 18, 87-123.

Douglass, H. R., & Spencer, P. L. Is it necessary to weight exercises in standard tests? Journal of Educational Psychology, 1923, 14, 109-112.

Dressel, P. L., & Schmid, J. Some modifications of the multiple-choice item. Educational and Psychological Measurement, 1953, 13, 574-595.

Flanagan, J. C. Factor analysis in the study of personality. Stanford: Stanford University Press, 1939.

Flanagan, J. C., & Davis, F. B. Table of correlation coefficients and item difficulty indices. Bronxville, N. Y.: Test Research Service, 1950.

Green, B. F., Jr. The sensitivity of Guttman weights. Paper presented at the meeting of the American Educational Research Association, Chicago, April, 1972.

Guilford, J. P. A simple scoring weight for test items and its reliability. Psychometrika, 1941, 6, 367-374.

Guilford, J. P. Psychometric methods. (2nd. ed.), New York: McGraw-Hill, 1954.

Guilford, J. P., Lovell, C., & Williams, R. Completely weighted versus unweighted scoring in an achievement examination. Educational and Psychological Measurement, 1942, 2, 15-21.

Gulliksen, H. Theory of mental tests. New York: Wiley, 1950.

Guttman, L. The quantification of a class of attributes: A theory and method of scale construction. In P. Horst, et. al., (Eds.), The prediction of personal adjustment. New York: Social Science Research Council, 1941.

Hawver, D. A. An experimental analysis of two partial information measures and acquiescent response bias. Unpublished doctoral dissertation, Temple University, 1969.

Hendrickson, G. The effect of differential option weighting on multiple-choice objective tests. Report No. 93, The Johns Hopkins University, 1971.

Holzinger, K. J. An analysis of the errors in mental measurement. Journal of Educational Psychology, 1923, 14, 278-288.

Kelley, T. L. The scoring of alternative responses with reference to some criterion. Journal of Educational Psychology, 1934, 25, 504-510.

Kelley, T. L. Fundamentals of statistics. Cambridge: Harvard University Press, 1947.

Kuder, G. F. A comparative study of some methods of developing occupational keys. Educational and Psychological Measurement, 1957, 17. 105-114.

McNemar, Q. Psychological Statistics. New York: Wiley, 1949.

Merwin, J. C. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. Journal of Educational Psychology, 1959, 50, 153-161.

Mosier, C. I. Problems and designs of cross-validation. Educational and Psychological Measurement, 1951, 11, 5-11.

Nedelsky, L. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19. (a)

Nedelsky, L. Ability to avoid gross error as a measure of achievement. Educational and Psychological Measurement, 1954, 14, 459-472. (b)

Odell, C. W. Further data concerning the effects of weighting exercises in new-type examinations. Journal of Educational Psychology, 1931, 22, 700-704.

Peatman, J. G. The influence of weighted true-false test scores on grades. Journal of Educational Psychology, 1930, 21, 143-147.

Potthoff, E. F., & Barnett, N. E. A comparison of marks based upon weighted and unweighted items in a new-type examination. Journal of Educational Psychology, 1932, 23, 92-98.

Powell, J. C. The interpretation of wrong answers from a multiple choice test. Educational and Psychological Measurement, 1968, 28, 403-412.

Reilly, R. R., & Jackson, R. Effects of item option weighting on validity and reliability of shortened forms of the GRE aptitude tests. Paper presented at the meeting of the American Educational Research Association, Chicago, April, 1972.

Sabers, D. L., & White, G. W. The effect of differential weighting of individual item responses on the predictive validity and reliability of an aptitude test. Journal of Educational Measurement, 1969, 6, 93-96.

Slakter, M. J. Risk taking on objective examinations. American Educational Research Journal, 1967, 4, 31-43.

Staffelbach, E. H. Weighting responses in true-false examinations. Journal of Educational Psychology, 1930, 21, 136-139.

Stalnaker, J. M. Weighting questions in the essay-type examination. Journal of Educational Psychology, 1938, 29, 481-490.

Stanley, J. D., & Wang, M. D. Differential weighting: A survey of methods and empirical studies. New York: College Entrance Examination Board, 1968.

Strong, E. K. Vocational interests of men and women. Stanford: Stanford University Press, 1943.

Swineford, F. Analysis of a personality trait. Journal of Educational Psychology, 1941, 32, 438-444.

Torgerson, W. S. Theory and methods of scaling. New York: Wiley, 1958.

West, P. V. The significance of weighted scores. Journal of Educational Psychology, 1924, 15, 302-308.

Wilks, S. S. Weighting systems for linear functions of correlated variables when there is no dependent variable. Psychometrika, 1938, 3, 23-40.